EPS
Escola Politècnica Superior
UdG

Projecte/Treball Fi de Carrera

Estudi: Enginyeria Informàtica. Pla 1997

Títol: ARI: Agent Recaptador d'Informació. Desenvolupament d'una aplicació que reculli inforació de portals webs dedicats a la gestió de premsa.

Document: Memòria del Projecte

Alumne: Alejandra Gómez Pérez

Director/Tutor: Gustavo Patow
Departament: Informàtica i Matemàtica Aplicada
Àrea: LSI

Convocatòria (mes/any): 09/07

1 Índex

1.1 Índex de continguts

1.2 Índex de taules

1.3 Índex de figures

2 Introducció i Objectius

2.1 Introducció

Avui en dia ens trobem davant d'una gran necessitat d'informació. Això és a causa de la gran explosió tecnològica que ha viscut el món en els últims anys. De fet, no fa gaire temps, la gent podia viure sense haver de tenir un mòbil o sense una connexió d'Internet d'alta velocitat... Aquestes noves necessitats, provocades per la societat a la que vivim, fan que hagin sorgit nous horitzons empresarials.

Així doncs, enfocant la visió cap a Internet es pot observar com han sorgit nous portals d'informació on els grans mitjans de comunicació espanyols han reflectit el que fins al moment feien en paper, radio o televisió. Fent un cop d'ull a aquests mitjans, han evolucionat cap l'apropament en tot moment de la informació més actual a l'usuari. En un primer moment, aquest portals es limitaven a transmetre el que ja havien donat prèviament en paper o per la ràdio o televisió, però clar, si quelcom aporta aquesta nova plataforma és la immensa capacitat per tenir la informació actualitzada al moment que es vulgui.

D'aquesta manera sorgeix el recull de notícies en RSS (Really Simple Syndication –RSS 2.0) que va atorgar als *webmasters* una eina per a publicar continguts d'una manera ràpida i senzilla. Aquest format per distribuir el contingut dels portals pels seus clients habituals va ser un gran esdeveniment, que va fer que les empreses ho veiessin com una nova font de consum i van passar de facilitar-los gratuïtament a fer que els usuaris s'haguessin de subscriure (i per tant, pagar quelcom) per poder rebre les notícies a les seves bústies de correu electrònic.

Veient aquesta evolució de la informació a Internet, va sorgir la idea d'aquest projecte, un motor de cerca orientat a la recaptació de notícies dispersades per les diferents pàgines on-line dels gran mitjans de comunicació espanyols, per tal que es pogués obtenir informació sobre "descriptors contractats"[1] pels usuaris del portal facilitat per aquest PFC.

Com a idea sobre el paper va semblar suficientment bona com per poder desenvolupar un Projecte Final de Carrera interessant i amb el gruix necessari d'investigació i innovació com per correspondre a una enginyeria superior.

En el seu origen, fa ja any i mig, es va començar a desenvolupar en una empresa on estava realitzant unes pràctiques de becària. Però degut a canvis a la legislació (referents a la Llei de Protecció de Dades i de la Propietat Intel·lectual), l'empresa en qüestió va decidir deixar de banda el projecte perquè no els resultava viable pels seus recursos de petita empresa. A pesar d'això, van decidir cedir el codi per poder continuar el desenvolupament del present Projecte Final de Carrera (PFC). Així doncs, encara que les expectatives empresarials amb les que va néixer el projecte no eren viables pel desenvolupament dins del marc de l'empresa en que s'estaven fent les pràctiques, es va canviar la visió original per una visió més acadèmica i d'investigació sobre el món dels portals de comunicació i els *spiders*[2] recaptadors d'informació.

Finalment, cal dir que el que es desenvoluparà a continuació és un projecte de caire investigador que vol esbrinar les possibilitats de recaptació d'informació que reporta un món tan extens i tant ple de possibilitats com és Internet, per portar a terme aquest PFC a continuació es desglossaran els objectius que es volen assolir i el corresponent anàlisi i disseny que el faran possible. L'estructura que es veurà, vindrà donada per la tècnica Extreme Programming, que donarà un caire d'evolució permanent al document, el qual pretén reflectir l'esmentat caire d'investigació del projecte.

[1] **Descriptors contractats:** expressió simbòlica, ja que al ser un PFC experimental no es realitza cap relació contractual amb els usuaris de la web dissenyada per aquest projecte.
[2] **Spider:** "aranya cercadora", motor de cerca on-line.

2.2 Objectius

Els objectius a assolir al final del desenvolupament d'aquest PFC són molt clars: l'obtenció d'un sèrie de notícies extretes de diferents pàgines web, dels diferents portals dels mitjans de comunicació espanyols, seguint uns paràmetres preestablerts.

Per poder arribar a desenvolupar la idea cal definir les fites a les quals es vol arribar. Tal com s'esmenta, l'objectiu principal està clar, obtenir notícies. Però, com arribar fins a elles? Com seleccionar les que es volen guardar al sistema? Per això calen uns passos previs d'investigació sobre el mode de procedir.

El primer objectiu és l'anàlisi de les necessitats que es volen cobrir per un hipotètic client de l'aplicació. D'aquesta manera s'estableix que es necessita un portal des d'on aquests clients es puguin registrar i accedir a una zona reservada per escollir sobre un llistat predeterminat uns mitjans a on es vol fer el seguiment i una altra secció a on es pot configurar el llistat de *descriptors*[3] que serviran per fer la selecció de notícies als mitjans escollits.

El segon objectiu és a l'àmbit algorítmic. Cal obtenir una metodologia de treball que permeti l'obtenció de la notícia. Per aconseguir això s'ha estudiat el següent:

- La llibreria *cUrl*[4]
- La utilització de patrons de cerca
- L'emmagatzemament a Base de Dades.

Aquest estudi permetrà obtenir de les pàgines web, mitjançant les funcionalitats de la llibreria cUrl, el contingut per poder fer el posterior anàlisi per poder determinar l'optimitat i decidir si cal o no guardar-ho al sistema. Per poder fer un bon anàlisi, s'aplicaran els conceptes extrets de la investigació sobre el món dels patrons i les expressions regulars[5], per finalment emmagatzemar-ho tot a la Base de Dades escollida.

Així doncs, els objectius a l'àmbit de la programació passen per tres etapes: descarregar les pàgines web necessàries, que es farà mitjançant les eines que proporciona la llibreria esmentada (cUrl). Amb aquesta eina es facilita la feina d'investigació d'obtenció de la informació base, des d'on es localitzaran els enllaços a les notícies que contenen les pàgines principals de cada portal.

Un cop es té descarregat el contingut de la pàgines, el següent objectiu és l'anàlisi. Hi ha tres tipus d'anàlisis:

1. Obtenir tots els enllaços que corresponen a notícies.
2. Filtrar els descriptors que es tenen a la Base de Dades per decidir si s'ha de guardar la notícia en qüestió o no
3. Un cop s'ha decidit si es necessita la notícia, analitzar la seva estructura interna, per guardar només les parts preestablertes (titular, entradeta i cos[6]) de la notícia.

Com a últim objectiu es troba la Base de Dades. Es necessari una estructura organitzada que permeti tenir tot totalment estructurat per poder obtenir les notícies i saber per quin o quins descriptors s'han escollit, de quin mitjà són o a quin o quins clients pertanyen.

[3] **Descriptor:** paraula o paraules que es cercaran dins de les notícies i que la seva ocurrència determinarà si aquesta s'emmagatzema al sistema o no.
[4] **cUrl:** llibreria PHP que permet la descàrrega del contingut d'una plana web a memòria.
[5] **Expressions regulars i patrons:** veure la secció 3.2.3 on s'explica el detall.
[6] **Titular, entradeta i cos:** veure la secció 3.1.1 on s'explica el detall de cada part de la notícia

2.3 Estructura documental

Al llarg de la documentació que s'aporta com a suport al projecte implementat es tractaran un seguit de punts claus que donaran forma al document.

Primerament, es troba la introducció i els objectius, que donen una visió global al lector de què es trobarà i què pot esperar de la lectura del document. Tot seguit, es troba un gran punt on s'expliquen tots els coneixements necessaris per poder desenvolupar el projecte i per poder entendre millor de què s'està parlant. De la mateixa manera, s'ofereix la planificació que s'ha seguit per realitzar el desenvolupament i la viabilitat que té tant a nivell empresarial com acadèmic.

Per un altre costat es troba estructurada la documentació en versions. Aquestes són les evolucions que ha anat patint el desenvolupament del projecte i a cada una es podrà veure el següent:

- L'anàlisi i el disseny que han requerit.
- Quines característiques especials tenen .
- Les diferències respecte les altres versions.
- Exemples de cerques
- Construcció i validació del sistema

Per últim, es plantegen les possibles ampliacions i millores que es poden donar tal com es deixa el projecte a l'última versió oferta. I al final de tot del document es localitzaran les conclusions extretes del procés de desenvolupament i els agraïments pertinents.

3 Coneixements previs

Per dur a terme aquest projecte, cal utilitzar uns coneixements que no són només els adquirits al llarg de la carrera. Per això, cal un aprenentatge previ sobre temes, tant del món periodístic (l'estructura d'una notícia, d'un portal web periodístic...), com informàtics (com llenguatges nous i ús de tecnologies diferents).

Per aquesta raó, en els apartats que es presentaran a continuació, es podrà veure el seguit de coneixements sobre les diferents àrees que es treballaran al llarg del desenvolupament del projecte.

3.1 Coneixements del medi

A aquest apartat el que es pretén donar es una visió general sobre els coneixements no tècnics que són requerits per poder desenvolupar el present projecte. Així doncs s'exposaran les bases més essencials, corresponents als coneixements sobre el medi sobre el qual es desenvolupa el PFC.

S'explicaran conceptes relacionats amb el món periodístic, breus referències al que s'ha de saber sobre què és una notícia, un RSS, el *press-clipping*[7]...

3.1.1 Notícia

És qualsevol informació d'un esdeveniment actual habitualment transmès pels mitjans de comunicació [1]: periòdic, televisió, ràdio o Internet. La notícia és informació, però el que la distingeix de qualsevol altre gènere és la seva condició de fet actual. A més la notícia està armada d'una forma especial: primer està l'epígraf, després està la baixada de títol i després el cos. Tota notícia respon a les següents preguntes: Quins? Què? Quan? Quant? On? Per què? Com?

Parts d'una notícia

- *Epígraf*: sol estar ubicat a la pàgina següent a la dedicatòria i anterior al pròleg.
- *Titular*: és l'element més extrem i visible de l'informe. Ha de referir-se al contingut i resumir-lo escaridament, fàcilment visible. Representa tant al seu contingut com al seu autor.
- *Baixada*: La baixada és l'encarregada d'aclarir el títol.
- *Cos de la notícia*: conté tot el contingut.
- *Copet, Lead o entradeta*: es presenta sempre sota el títol. És una amplificació d'aquest i consisteix en una síntesis de la informació, amb dades precisos sobre aquesta.
- *Fotografia del tema*: opcional.
- *La crònica*: és una informació interpretada sobre fets actuals on es narren un succés passat que es relaciona amb un actual; en altres paraules, maneja i juga amb el temps.

Característiques principals

- *Veracitat:* els fets o successos han de ser veritables i, per tant, verificables.
- *Objectivitat:* el periodista no s'ha de veure reflectit en ella mitjançant la introducció de cap opinió o judici de valor. A la notícia no ha d'aparèixer qui l'ha redactat, només s'endevinarà que té un autor perquè en ella es dóna una selecció de la realitat, de manera que el periodista escull els elements que li semblen interessants i rellevants. Però en cap cas es mostrarà la seva opinió.
- *Claredat:* els fets han de ser exposats de manera ordenada i lògica.
- *Brevetat:* els fets han de ser presentats breument, sense reiteracions o dades irrellevants.
- *Generalitat:* la notícia ha de ser d'interès social i no particular.
- *Actualitat:* els fets han de ser actuals o recents.

[7] **Press-clipping:** recopilacions comercials d'articles de premsa, concepte ampliat al punt 3.1.3

- *Novetat:* els successos han de ser nous, desacostumats i rars.
- *Interès humà:* la notícia ha de ser capaç de produir una resposta afectiva o emocional als receptors.
- *Proximitat:* els successos entregats provoquen major interès si són propers al receptor.
- *Prominència:* la notícia provoca major interès si les persones involucrades són importants i conegudes.
- *Conseqüència:* té interès noticiós tot el que afecti a al vida de les persones.
- *Oportunitat:* mentre més ràpid es doni a conèixer un fet noticiós major valor posseeix.
- *Desenllaç:* algunes notícies mantenen el interès del públic a espera del desenllaç que pugui resultar sorprenent.
- *Tema:* les notícies relacionades amb certs àmbits del quefer humà resulten atractives en sí mateixes: avanços científics.
- *Servei:* una notícia es pot percebre com a tal en funció del servei que presti. Que ajudi a prendre decisions.

3.1.2 RSS

Forma part de la família dels formats XML desenvolupat específicament per tot tipus de llocs que s'actualitzen amb freqüència i per mitjà del qual es pot compartir la informació i utilitzar-la en altres llocs web o programes [2]. Això es coneix com a redifusió o sindicació web.

El RSS no és una altre cosa que un senzill format de dades que és utilitzat per sindicar (redifondre) continguts a subscriptors d'un lloc web. El format permet distribuir contingut sense necessitat d'un navegador, el qual també es pot veure com desavantatge ja que necessita de la instal·lació d'un altre software. Alguns avanços han permès utilitzar el mateix navegador per veure els continguts RSS mitjançant programació dels denominats scripts d'interpretació. Així també les noves versions dels navegadors permetran llegir els RSS sense necessitat de software addicional. L'acrònim s'utilitza per els següents estàndards:

- Rich Site Summary (RSS 0.91)
- RDF Site Summary (RSS 0.9 i 1.0)
- Really Simple Syndication (RSS 2.0)

Els programes que llegeixen i presenten fonts RSS de diferents procedències s'anomenen agregadors.

Gràcies als agregadors o lectors de *feeds*[8] es poden obtenir resums de tots els llocs que es desitgi des de l'escriptori del sistema operatiu, programes de correu electrònic o per mitjà d'aplicacions web que funcionen com agregadores. No és necessari obrir el navegador i visitar dotzenes de webs.

Però el que és veritablement important és que a partir d'aquest format s'està desenvolupant una cadena de valor nova al sector dels continguts que està canviant les formes de relació amb la informació tant de professionals i empreses del sector com dels usuaris. Varies empreses estan explorant noves formes d'ús i distribució de la informació.

La sindicació web no és només un fenomen vinculat als *weblogs*[9], encara que han ajudat molt a la seva popularització. Sempre s'han sindicat continguts i s'ha compartit tot tipus d'informació en format XML. D'aquesta manera es pot oferir continguts propis para que siguin mostrats en altres pàgines web de forma integrada, el que augmenta el valor de la pàgina que mostra el contingut i també ens genera més valor, ja que normalment la sindicació web sempre enllaça amb els continguts originals.

[8] **Feeds:** programes o llocs que permeten llegar fonts web
[9] **Weblogs:** lloc web que periòdicament actualitzat que recopila cronològicament text o articles d'un o varis autors.

3.1.3 Press-Clipping

Servei que entra a l'àmbit de les relacions públiques. Consisteix en detectar tot el que es difon als mitjans de comunicació sobre un tema o una organització determinats, categoritzant la informació i recopilant-la de tal manera que pugui ser distribuïda fàcilment [3].

És habitual que el seguiment de mitjans ho porti a terme empreses especialitzades que són contractades per organitzacions o per professionals de les relacions públiques. D'aquesta manera, tradicionalment, aquestes empreses obtenien la premsa escrita als grans mitjans de comunicació i ho analitzaven amb un equip de gent que es dedicava al escaneig i posterior lectura de les notícies per poder classificar-la segons les necessitat. De tal manera, el naixement dels portals d'aquests mateixos mitjans va crear unes noves expectatives que es van veure frustrades per les queixes d'aquest mateixos portals i la redacció de noves lleis (aquest punt s'amplia a l'apartat *3.4. Obstacles Legals*).

3.2 Coneixements tècnics

Relacionat amb temes més tècnics, es va trobar la necessitat d'investigar un parell de qüestions. La primera d'aquestes no s'ha donat al llarg de la carrera, tot el tema corresponent a la llibreria cUrl de PHP, per l'obtenció del contingut de les pàgines web que es volen analitzar. També, referent al món d'Internet i la programació de pàgines web, cal una visió més profunda del llenguatge HTML, amb aquest llenguatge s'ha pogut treballar al llarg de la carrera si s'han fet optatives com Multimèdia o NTSI[10], assignatures les dos de tercer d'ETIG. I per un altra costat, per poder analitzar aquest contingut, es va veure que els coneixements adquirits a assignatures com TALLF[11], s'haurien d'ampliar en alguns aspectes, com en les tècniques de desenvolupament d'expressions regulars.

3.2.1 Llibreria cUrl

cUrl és una llibreria de funcions per connectar amb servidors i treballar amb ells. Aquest treball es realitza amb format URL [4]. Es a dir, serveix per a realitzar accions sobre arxius que hi ha a URLs d'Internet, suportant els protocols més comuns, com http, ftp, https, etc.

Pel que fa a PHP, cUrl està integrat dintre, de manera que aquestes llibreries també es poden utilitzar des d'scripts PHP. Encara que per això PHP s'ha d'haver instal·lat amb suport a cUrl i no és així en tots els casos.

Així doncs PHP, que és el idioma escollit per desenvolupar el projecte, suporta *libcurl*, una llibreria creada per Danile Stenberg, que permet la connexió i comunicació amb varis tipus de servidors diferents i amb molts tipus de protocols diferents. Actualment, *libcurl* suporta els protocols http, https, ftp, gopher, telnet, dict, file i ldap. A més, *libcurl* també suporta certificats HTTPS, mètodes HTTP POST i HTTP PUT, enviament per FTP, enviament mitjançant HTTP d'arxius en formularis HTML, servidors proxy, cookies i autenticació usuari+contrasenya.

Requeriments: per poder utilitzar les funcions cUrl, s'haurà d'instal·lar el paquet CURL. PHP requereix que s'utilitzi CURL 7.0.2-beta o superior. Des de la versió 4.2.3 de PHP es necessita, al menys, CURL 7.9.0 o superior. A partir de la versió 4.3.0, es necessita una versió de CURL 7.9.8 o superior i a partir de la versió PHP 5.0.0, la versió de CURL necessària ha de ser superior a 7.10.5.

[10] **NTSI**: Noves Tecnologies de Sistemes d'Informació (3er d'ETIG)
[11] **TALLF**: Teoria d'Autòmats i Llenguatges Formals (de 4rt d'E.INF)

3.2.2 HTML

HTML (Acrònim de *Hyper Text Markup Language*, en català, "llenguatge de marcat d'hipertext"), és un llenguatge de marcat que deriva de l'SGML dissenyat per estructurar textos i relacionar-los en forma d'hipertext. Gràcies a Internet i als navegadors web, s'ha convertit en un dels formats més populars que existeixen per a la construcció de documents [5].

Història de l'HTML

En el seu origen, l'HTML era un llenguatge dissenyat per compartir informació científica entre científics de tot el món. Era purament un llenguatge estructural, en què no hi havia forma de descriure l'aparença de les pàgines (ni tan sols la possibilitat de posar un text en negreta o cursiva). Més endavant s'hi van afegir nombroses opcions per formatar i presentar text i gràfics.

A mitjans de la dècada de 1990 van començar les ampliacions de l'HTML per aconseguir la presentació desitjada, però sempre des de diferents perspectives de diferents desenvolupadors, que van acabar amb diverses solucions no estàndards per a diferents navegadors. Això va provocar l'aparició d'un consorci que controla l'evolució de l'HTML: el W3C (World Wide Web Consortium).

Aquesta evolució tenia un punt clau: la separació del contingut i l'aparença. Amb la versió 4 de l'HTML es recomanava un altre mecanisme per controlar la visualització del nostre contingut HTML: els fulls d'estil (CSS: Cascading Style Sheets).

El W3C recomana l'ús de l'XHTML, que manté la mateixa sintaxi i els mateixos mecanismes que l'HTML, però reformulat amb les normes d'un XML, preparant-se així per aprofitar les avantatges d'aquest llenguatge.

Per altre banda el WHATWG —grup de treball composat per la Fundació Mozilla i Opera Inc.— estan plantejant una especificació per un HTML 5 estenent l'HTML 4.01 i el <abbr title="Document Object Model">DOM</abbr>. L'**HTML 5** intenta millorar la part d'aplicació web amb l'especificació Web Forms 2.0. Aquest grup surt com a reacció pel canvi brusc del pas d'HTML a XHTML que, si no fos per l'Apèndix C de l'especificació XHTML 1.0 no es podria usar en navegadors que no suporten el MIME type `application/xhtml+xml`.

No s'ha d'entendre el WHATWG com una organització paral·lela al W3C sinó un grup complementari ja que quan té un esborrany el proposa al W3C per tal d'estandarditzar-lo.

La darrera especificació vigent és l'XHTML 1.1 que ja no contempla cap compatibilitat amb versions anteriors i, per tant, només es pot servir com a `application/xhtml+xml` excloent qualsevol navegador antic.

El punt més polèmic actualment és la proposta d'especificació (en estat d'esborrany) XHTML 2.0 que deixa de ser compatible amb versions anteriors no només a nivell de MIME type sinó que l'estructura de document i elements estructurals canvien.

Etiquetes bàsiques

Les etiquetes bàsiques d'HTML, d'obligada presència en tot document són:

- `<!DOCTYPE>`: És l'etiqueta que permet definir el tipus de document HTML que s'empra. Existeixen tres tipus bàsics: l'estricte (Strict), el transicional (Transitional) i el de marcs (Frameset).
- `<html>`: És l'etiqueta arrel de qualsevol document HTML o XHTML.
- `<head>`: Defineix la capçalera del document HTML. Permet declarar metainformació del document que no es mostra directament en el navegador. Aquesta informació és d'especial rellevància pels indexadors i cercadors automàtics.

- `<body>`: Defineix el cos del document. Aquesta és la part del document HTML que es mostra en el navegador.

Dintre de la capçalera `<HEAD>` hi podem trobar:

- `<title>`: Permet definir el títol de la pàgina. En navegadors gràfics el contingut del `title` apareix a la barra del títol a sobre de la finestra.
- `<meta>`: Permet definir metainformacions del document tals com l'autor, la data de realització, la codificació del document (UTF, ISO, etc.), les paraules clau i la descripció del mateix
- `<LINK>`: Permet definir metadades complementàries a les del `meta` tals com el document anterior, el següent, el capítol al qual pertany el document, la pàgina glossari, etc.

Dintre del cos `<BODY>` hi podem trobar:

- `<a>`: Etiqueta ancla. Crea un enllaç a un altre document o a una altra zona del mateix, segons els atributs.
- `<h1>, <h2>,... <h6>`: capçaleres o títols del document, acostumen a distingir-se per mida.
- `<div>`: Divisió estructural de la pàgina.
- `<p>`: Paràgraf.
- `<br>`: Salt de línia.
- `<table>`: Indica el començament d'una taula, després s'haurà de definir les files amb `<tr>` i les cel·les dintre de les files amb `<td>`.
- `<ul>`: Llista desordenada (sense numerar). Els ítems es defineixen amb `<li>`.
- `<ol>`: Llista ordenada (numerat). Els ítems es defineixen amb `<li>`.
- `<dl>`: Llista de definició. Hi ha dos tipus d'ítem; el `dt` i el `dd`.
- `<dt>`: Terme a definir.
- `<dd>`: Definició del terme.

Excepte unes poques etiquetes, la majoria requereixen ser tancades escrivint la mateixa etiqueta precedida d'una barra "/".

Exemples:

- `<html>...</html>`
- `<table><tr><td>Contingut d'una cel·la</td></tr></table>`
- `<script>Codi d'un script encastat en la pàgina</script>`

Nocions bàsiques d'HTML

El llenguatge HTML pot ser creat i editat amb qualsevol editor de text bàsic, com pot ser Gedit, el *Bloc de Notes* de Windows, o qualsevol altre editor que admeti test sense format com GNU Emacs, Mircrosoft Wordpad, TextPAd, Vim, etc.

Existeixen a més, altres programes per la realització de llocs Web o edició de codi HTML, com per exemple Microsoft FrontPage, el qual té un format bàsic semblant a la resta de programes d'Office. També existeix el famós software de Macroedia (que va adquirir Adobe) anomentat Dreamweaver, essent un dels més utilitzats a l'àmbit del disseny i programació Web. Aquest programes se'ls

coneix com editors WYSIWIG[12]. Això significa que són editors en els quals es veu el resultat del que s'està editant en temps real a mida que es va desenvolupant el document. Ara bé, això no significa una manera diferent de realitzar llocs web, sinó que és una forma un tant més simple ja que aquest programes, a més de tenir la opció de treballar amb la vista preliminar, tenen la seva pròpia secció HTML la qual va generant tot el codi a mida que es va treballant.

Combinar aquest dos mètodes resulta molt interessant, ja que d'alguna manera s'ajuden entre sí. Per exemple; si s'edita tot en HTML i de sobte s'oblida un codi o etiqueta, simplement s'ha de dirigir a l'editor visual i es continua allà l'edició, o viceversa, ja que hi ha casos en que surt més ràpid i fàcil escriure directament el codi d'alguna característica que es vulgui adherir-li al lloc, que buscar l'opció al programa mateix.

HTML utilitza etiquetes o marques, que consisteixen en breus instruccions de començament i final, mitjançant les quals es determina la forma en la que ha d'aparèixer en el navegador el text, així com també les imatges i els demés elements, a la pantalla de l'ordinador.

Tota etiqueta s'identifica perquè està tancada entre els signes de *menor que* i *major que* (<>), i algunes tenen atributs que poden prendre algun valor. En general les etiquetes s'aplicaran de dues formes especials:

- S'obren i es tanquen, com per exemple: negreta que es veuria en el navegador com **negreta**.
- No poden obrir-se i tancar-se, com <hr> que es veuria en el navegador com una línia horitzontal.
- Altres que poden obrir-se i tancar-se, com <p>
- Les etiquetes bàsiques o mínimes són:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" http://www.w3.org/TR/html4/strict.dtd">
        <html lang="es">
            <head>
                <title>Exemple</title>
            </head>
            <body>
                <p>Exemple</p>
            </body>
        </html>
```

Per conèixer una mica més

Es pot provar a escollir *Veure codi font* al navegador, així es podrà veure la informació que realment s'està rebent al navegador i com l'està interpretant. Per exemple, a Internet Explorer o a Firefox simplement s'ha de pressionar *Veure* al menú principal, després s'ha de pressionar *Codi font* i ja estarà, ja es pot veure el codi font de la pàgina que s'està veient a l'explorador o també una altra forma més senzilla es fent *click* al botó dret del *mouse* i seleccionar l'opció *Veure codi font de la pàgina*.

3.2.3 Expressions regulars

Una expressió regular, també coneguda com a patró, és una expressió que descriu un conjunt de cadenes sense enumerar els seus elements [6]. Per exemple, el grup format per les cadenes *Handel*, *Händel* i *Heandel* es descriu mitjançant el patró "H(a|ä|ae)ndel". La majoria de les formalitzacions proporcionen els següents constructors: una expressió regular és una forma de representar als llenguatges regulars (finits o infinits) i es construeix utilitzant caràcters de l'alfabet sobre el qual es defineix el llenguatge. Específicament, les expressions regulars es construeixen utilitzant els operadors: concatenació, unió i clàusula de Kleene.

[12] WYSIWYG: *What You See Is What You Get*, en català, "el que veus és el que obtens"

- La concatenació consisteix en unir dues paraules d'un determinat llenguatge. Cal posar els símbols de la segona paraula darrera dels de la primera. Així doncs, es equivalent fer: casa·blanca com casablanca.
- La unió dóna com a resultat els elements comuns de les paraules que es volen unir.
- La clàusula de Kleene, també anomenada estrella Kleene o tancament estrella és una operació unària que s'aplica sobre un conjunt de cadenes de caràcters o un conjunt de símbols o caràcters, i representa el conjunt de les cadenes que es poden formar prenent qualsevol número de cadenes del conjunt inicial, possiblement repeticions i concatenant-les entre si. La seva aplicació a un conjunt *V* es simbolitza com *V**.

A continuació es presentaran més operadors que s'utilitzen en la construcció de les expressions:

Alternació |: Una barra vertical separa les alternatives. Per exemple: "casa|llar" definirà tant les cadenes *casa* com *llar* .

Quantificació: Un quantificador després d'un caràcter especifica la freqüència amb la que aquest pot aparèixer. Els quantificadors més comuns són +,? i *:

- **+:** el signe més indica que el caràcter al que segueix ha d'aparèixer al menys una vegada. Per exemple, "ho+la" descriu el conjunt infinit *hola*, *hoola,hooola, hoooola*, ...
- **?:** el signe d'interrogació indica que el caràcter al que segueix pot apareixer com a molt una vegada. Per exemple, "ob?scur" por reflexar-se amb *oscur* i *obscur*.
- ***:** l'asterisc indica que el caràcter al que segueix pot aparèixer cero, una o més vegades. Per exemple, "0*42" casa amb 42, 042, 0042, 00042,...

*Agrupació***:** Els parèntesis es poden utilitzar per definir l'àmbit i precedència dels demès operadors. Per exemple, "(p|m)are" reconeix el mateix que "pare|mare" i "(des)?amor" coincideix amb *amor* i *desamor*.

Els constructors poden combinar-se lliurament dins de la mateixa expressió, pel que "H(ae?|ä)ndel" equival a "H(a|ae|ä)ndel".

La sintaxis precisa de les expressions regulars canvia segons les eines i aplicacions considerades, i es descriuen amb més detall a continuació.

La seva utilitat més obvia és la de descriure un conjunt de cadenes, el que resulta d'utilitat en editors de text i aplicacions per buscar i manipular textos. Molts llenguatges de programació admeten l'ús d'expressions regulars amb aquest fi. Per exemple, *Perl* té un potent motor d'expressions regulars directament inclòs a la seva sintaxis. Les eines proporcionades per les distribucions de Unix van ser les primeres en popularitzar el concepte d'expressió regular.

Respecte a aplicacions, nombrosos editors de text i altres utilitats (especialment en el sistema operatiu Unix), com per exemple *sed* i *awk*[13], utilitzen expressions regulars per, per exemple, buscar paraules al text i canviar-les per alguna altra cadena de caràcters.

A la programació, les expressions són un mètode per mitjà del qual es poden realitzar cerques dins de cadenes de caràcters. Sense importar si la cerca requerida és de dos caràcters en una cadena de 10 o si es necessari trobar totes les coincidències d'un patró definit de caràcters a una arxiu de milions de caràcters, les expressions regulars proporcionen una solució pel problema. Addicionalment, un ús derivat de la cerca de patrons és la validació d'un format específic a una cadena de caràcters donada, com per exemple dates o identificadors.

[13] **SED:** és un editor de flux, una potent eina de tractament de text pel sistema operatiu Unix que acepta com entrada un arxiu, ho llegeix i modifica línea a línea mostrant el resultat per pantalla. **AWK**: llenguatge de programació dissenyat per a processar dades basades en text, ja siguin fitxers o fluxes de dades.

Per poder utilitzar les expressions regulars al programar es necessari tenir accés a un motor de cerca amb la capacitat d'utilitzar-les. Es possible classificar els motors disponibles en dos tipus: motors pel programador i motors per l'usuari final.

- *Motors per l'usuari final:* són programes que permeten realitzar cerques sobre el contingut d'un arxiu o sobre un text extret i col·locat al programa. Estan dissenyats per permetre a l'usuari realitzar cerques avançades utilitzant aquest mecanisme, no obstant és necessari aprendre a redactar expressions regular adequades per poder utilitzar-los eficientment. Aquest són alguns dels programes disponibles: *grep*, programa dels sistemes operatius Unix i Linux; *PowerGrep*,versió de *grep* pels sistemes operatius Windows; *RegexBuddy_* ajuda a crear expressions regulars de forma interactiva i després permet a l'usuari utilitzar-les i guardar-les; *EditPad Pro*, permeten realitzar cerques amb expressions regulars sobre arxius i les mostra per mitjà de codi de colors per facilitar la seva lectura i comprensió.
- *Motors pel programador:* permeten automatitzar el procés de cerca de manera que sigui possible utilitzar-ho moltes vegades per un propòsit específic. Aquestes són algunes de les eines de programació disponibles que ofereixen motors de cerca amb suport a expressions regulars: *Java*, existeixen llibreries fetes per java que permeten l'ús de RegEx , i Sun planeja incorporar-ho al SDK; *JavaScript*, a partir d ela versió 1.2 (ie4+, ns4+) JavaScript té suport integrat per expressions regulars, el que significa que les validacions que es realitzen normalment a una pàgina web es podrien simplificar molt si el programador sabés utilitzar aquesta eina; *Perl*, és el llenguatge que va fer créixer a les expressions regulars a l'àmbit de la programació fins arribar al que són avui en día; *PCRE*, llibreria de ExReg per C,C++ i altres llenguatges que poden utilitzar llibreries dll (Visual Basic 6, per exemple); *PHP*, tenen dos tipus diferents d'expressions regulars disponibles pel programador; *Python*, llenguatge de "scripting" popular amb suports a expressions regulars; i *.Net Framework*, proveeix un conjunt de classes mitjantçant les quals es possible utilitzar expressions regulars per fer cerques, reemplaçar cadenes i validar patrons.

Les expressions regulars permeten trobar porcions específiques de text dins d'una cadena més gran de caràcters. Així, si es necessari trobar el text "lot" a l'expressió " l'ocellot té un lot de peces" qualsevol motor de cerca seria capaç d'efectuar aquesta tasca. No obstant, la majoria dels motors de cerca permeten addicionalment especificar que es desitja trobar només paraules completes, solucionant aquest problema. Les expressions regulars permeten especificar totes aquestes opcions addicionals i moltes altres sense necessitat de configurar opcions especials, sinó utilitzant el mateix text de cerca com un llenguatge que permet enviar-li al motor de cerca exactament el que es desitja trobar en tots els casos, sense necessitat d'activar opcions addicionals al realitzar la cerca.

Per especificar opcions dins del text a buscar, s'utilitza un llenguatge o convenció mitjançant el qual se li transmet al motor de cerca el resultat que es desitja obtenir. Aquest llenguatge li dona un significat especial a una sèrie de caràcters. Per tant, quan el motor de cerca d'expressions regulars trobi aquest caràcters no els trobarà al text de forma literal, sinó que buscarà el que els caràcters signifiquen. A aquest caràcters se'ls anomena algunes vegades "meta-caràcters". Ara es presentarà un llistat dels principals, la seva funció i com els interpreta el motor d'expressions regulars.

- *El punt:* El punt es interpretat pel motor de cerca com qualsevol altre caràcter exceptuant els caràcters que representen un salt de línia, a menys que se l'indiqui al motor d'expressions regulars el contrari. Per tant, si aquesta opció se li diu al motor de cerca que s'utilitzi, el punt li dirà al motor que trobi qualsevol caràcter incloent els salts de línia. A l'eina *EditPad Pro* això es fa per mitjà de l'opció "punt correspon a nova línia" a les opcions de cerca.

  El punt s'utilitza de la següent forma: si se'l diu al motor de *RegEx* que busqui "g.t" a la cadena "el gat de pedra a la gòtica porta de getisboro goot" el motor de cerca trobarà "gat", "gòt" i per últim "get". Cal fixar-se que el motor de cerca no troba "goot"; això és perquè el punt representa un sol caràcter i únicament un. Si es necessari que el motor troba també la expressió "goot", serà necessari utilitzar repeticions, les quals són explicades més endavant.

Encara el punt és molt útil per trobar caràcters que no es coneixen, és necessari recordar que correspon a qualsevol caràcter i que moltes vegades això no és el que es requereix. Cal tenir en compte que és molt diferent indicar-li al motor de cerca que busqui qualsevol caràcter a dir-li que buqui qualsevol caràcter alfanumèric, que és mes restrictiu o passar a especificar que busqui qualsevol dígit o qualsevol no-dígit o qualsevol no-alfanumèric, ja que s'estan deixant fora caràcters que el punt per sí mateix inclou. Per aquesta raó cal anar amb compte abans d'utilitzar el punt i obtenir resultats no desitjats.

- *La barra inversa o contrabarra "\":* s'utilitza per "marcar" el següent caràcter de l'expressió de cerca de manera que aquest adquireix un significat especial o deixi de tenir-lo. O sigui, la barra inversa no s'utilitza mai per sí sola, sinó en combinació amb altres caràcters. Al utilitzar-lo per exemple en combinació amb el punt "\." aquest deixa de tenir el seu significat normal i es comporta com un caràcter literal.

  De la mateixa manera, quan es col·loca la barra inversa seguida de qualsevol dels caràcters especials que discutirem a continuació, aquest deixen de tenir el seu significat especial i es converteixen en caràcters de cerca literal.

  Com ja es va mencionar amb anterioritat, la barra inversa també pot donar-se significat especial a caràcters que no ho tenen. A continuació una llista amb alguns exemples explicats:

  - **\t:** representa un tabulador
  - **\r:** representa el "retorn al inici" o sigui el lloc en que la línia torna a iniciar
  - **\n:** representa la "nova línia" el caràcter per mitjà del qual una línia da inici. Es necessari recordar que en Windows es necessària una combinació de \r\n per començar una nova línia, mentre que en Unix solament s'utilitza \n.
  - **\d:** representa un dígit.
  - **\w:** representa qualsevol caràcter alfanumèric.
  - ...

- *Els claudàtors "[]":* la funció dels claudàtors en el llenguatge de les expressions regulars és representar "classes de caràcters", o sigui, agrupar caràcters en grups o classes. Són útils quan es necessari buscar un caràcter d'un grup de caràcters. Dins dels claudàtors és possible utilitzar el guió "-" per especificar rangs de caràcters. Addicionalment, els metacaràcters perden el seu significat i es converteixen en literals quan es troben dins dels claudàtors. Per exemple, el punt representa un caràcter literal i no un metacaràcter, pel que no es necessari antecedir-lo a la barra inversa. L'únic caràcter que es necessari antecedir amb la barra inversa dins dels claudàtors és la pròpia barra inversa "\\".

- *La barra "|":* serveix per indicar una de varies opcions. Per exemple, l'expressió regular "a|e" trobarà qualsevol "a" o "e" dins del text. L'expressió regular "est|nord|sud|oest" permetrà trobar qualsevol dels noms dels punts cardinals. La barra s'utilitza comunament en conjunt amb altres caràcters especials.

- *El signe de dòlar "$":* representa el final de la cadena de caràcters o el final de la línia, si s'utilitza en mode multi-línia. No representa un caràcter en especial sinó una posició. Si s'utilitza la expressió regular "\.$" el motor trobarà tots els llocs on un punt finalitzi la línia, el que es útil per avançar entre paràgrafs.

- *L'accent circumflex "^":* Aquest caràcter té una doble funcionalitat, que difereix quan s'utilitza individualment i quan s'utilitza en conjunt amb altres caràcters especials. En primer lloc la seva funcionalitat com caràcter individual: el caràcter "^" representa el inici de la cadena. Per lo tant, si s'utilitza l'expressió regular "^[a-z]" el motor trobarà tots els paràgrafs que den començament amb una lletra minúscula. Quan s'utilitza en conjunt amb els claudàtors de la següent manera "[^\w]" permet trobar qualsevol caràcter que NO es trobi dins del grup indicat.

La utilització en conjunt dels caràcters especials "^" i "$" permet realitzar validacions de forma senzilla. Per exemple "\d$" permet assegurar que la cadena a verificar representa un únic dígit.

- *Els parèntesis "()":* de manera similar que els claudàtors, les parèntesis serveixen per agrupar caràcters, no obstant existeixen diverses diferències fonamentals entre els grups establerts per mitjà de claudàtors i els grups establers per parèntesis:
    - Els caràcters especials conserver el seus significat dins dels parèntesis
    - Els grups establers amb parèntesis estableixen una "etiqueta" o "punt de referència" pel motor de cerca que pot ser utilitzada posteriorment.
    - Utilitzats en conjunt amb la barra "\" permet fer cerques opcionals.
    - Utilitzat en conjunt amb altes caràcters especials ofereix funcionalitats addicionals esmentades als punts següents.
- *El signe de pregunta "?":* el signe de pregunta té diverses funcions dins del llenguatge de les expressions regulars. La primera d'elles és especificar que una part de la cerca és opcional. Per exemple, l'expressió regular "ob?scuritat" permet trobar tan "oscuritat" com "obscuritat". En conjunt amb els parèntesis rodons permet especificar que un conjunt major de caràcters és opcional; per exemple, "Nov(\.|iembre|embre)?" permet trobar tant "Nov" com "Nov.", "Noviembre" i "Novembre". Com es va mencionar anteriorment, els parèntesis permeten establir un "punt de referència" pel motor de cerca.
- *Claus "{}":* Comunament les claus són caràcters literals quan s'utilitzen per separat en una expressió regular. Per que adquireixin la seva funció de metacaràcters es necessari que tanquin un o diversos números separats per coma i que estiguin col·locats a la dreta d'una altre expressió regular de la següent forma: "\d{2}". Aquesta expressió li diu al motor de cerca que trobi dos dígits contigus. Utilitzant aquesta fórmula es podria convertir l'exemple "^\d\d/\d\d/\d\d\d\d$" que serveix per validar un format de data en "^\d{2}/\d{2}/\d{4}$" per una major claredat en la lectura de l'expressió.
- *L'asterisc "*":* serveix per trobar quelcom que es trobi repetit 0 o més vegades. Per exemple, utilitzant l'expressió "[a-zA-Z]\d*" serà possible trobar tant "H" com "H1", "H01", "H100"... és a dir, una lletra seguida d'un número indefinit de dígits. Es necessari tenir cura amb el comportament de l'asterisc, ja que aquest per defecte tracta de trobar la major quantitat possible de caràcters que corresponguin amb el patró que es busca. D'aquesta forma si s'utilitza "\(.*\)" per trobar qualsevol cadena que es trobi entre parèntesis i s'aplica sobre el text "Veure (Fig. 1) i (Fig. 2)" s'esperaria que el motor de cerca trobés els text "(Fig. 1)" i "(Fig. 2)", no obstant, degut a aquesta característica, en comptes d'això, en el seu lloc trobarà el text "(Fig. 1) i (Fig. 2)". Això succeeix perquè l'asterisc li diu al motor de cerca que ompli tots els espais possibles entre dos parèntesis. Per obtenir el resultat desitjat s'ha d'utilitzar l'asterisc en conjunt amb el signe de pregunta de la següent manera: "\(.*?\)". Això es equivalent a dir-li al motor de cerca que "Troba un parèntesis d'obertura i després troba qualsevol caràcter repetit fins que trobi un parèntesis de tancament".
- *El signe de suma "+":* s'utilitza per trobar una cadena que es trobi repetida 1 o més vegades. A diferència de l'asterisc, l'expressió "[a-zA-Z]\d+" trobarà "H1" però no trobarà "H". També es possible utilitzar aquest metacaràcter en conjunt amb el signe de pregunta per limitar fins a on s'efectua la repetició.
- *Grups anònims:* s'estableixen cada vegada que es tanca una expressió regular entre parèntesis, pel que l'expressió "([a-zA-Z]\w*?)" defineix un grup anònim que tindrà com a resultat que el motor de cerca emmagatzemarà una referència al text que correspongui a l'expressió tancada entre els parèntesis.

Com es poden utilitzar els grups anònims que s'estableixen? La forma més immediata d'utilitzar aquests grups que es defineixen és dins de la mateixa expressió regular. Aquests grups es porten a terme utilitzant la barra inversa "\" seguida del número del grup al que es desitja fer referència (a

mode d'etiqueta), de la següent manera: "<([a-zA-Z]\w*?)>".*?</\1>". Aquesta expressió regular ajudarà a trobar tant la cadena "Aquesta" com la cadena "proba" en el text "Aquesta és una proba" a pesar de que la expressió no conté els literals "font" i "b".

3.3 Decisions respecte l'entorn de treball

Per realitzar aquest projecte s'han necessitat de varies eines, tant a nivell de programació com de Base de Dades. A continuació s'explicarà la Base de Dades emprada (MySQL) i el llenguatge de programació escollit (PHP) per dur a terme la realització d'aquest PFC.

3.3.1 MySQL

És un sistema de gestió de Base de Dades relacional, multifil i multiusuari amb més de sis milions d'instal·lacions [7]. MySQL AB desenvolupa MySQL com software lliure en un esquema de llicencies dual. Per un costat ho ofereix sota *GNY GLP*[14], però empreses que vulguin incorporar-ho en productes privatius poden comprar a l'empresa una llicència que els permeti aquest ús.

Està desenvolupat en la seva gran part en ANSI C[15]

Al contrari de projectes com *Apache*, on el software es desenvolupat per una comunitat pública, i el dret de còpia del codi està en poder de l'autor individual, MySQL és propietat i està patrocinat per una empresa privada, que posseeix el dret de còpia de la major part del codi. Això és el que possibilita l'esquema de llicències abans esmentat. A més de la venta de llicencies privatives, la companyia ofereix suport i serveis. Per les seves operacions contracten treballadors al voltant del món que col·laboren via Internet. MySQL AB va ser fundat per David Axmark, Allan Larson, i Michael Widenius.

Historia del projecte

SQL (Llenguatge de Consulta Estructurat) va ser comercialitzat per primera vegada al 1981 per IBM, el qual va ser presentat en ANSI i des de llavors ha sigut considerat com un estàndard per les bases de dades relacionals. Des de 1986, l'estàndard SQL ha aparegut en diferents versions com per exemple: SQL:92, SQL:99, SQL:2003. MySQL és una idea originaria de l'empresa (de codi obert) MySQL AB establerta inicialment a Suècia al 1995 i els fundadors de la qual són, els esmentats prèviament, David Axmark, Allan Larsson, i Michael "Monty" Widenius. L'objectiu que persegueix aquesta empresa consisteix en que MySQL compleixi l'estàndard SQL, però sense sacrificar velocitat, fiabilitat o usabilitat.

Michael Widenius a la dècada dels 90 va tractar d'utilitzar un llenguatge anomenat mSQL per connectar les taules utilitzant rutines de baix nivell ISAM[16], no obstant, mSQL no era ràpid i flexible per les seves necessitats de tractament de les taules i les rutines que es volien executar. Això va portar a crear una API SQL denominada MySQL per bases de dades molt similar a la de mSQL però més portable.

La procedència de nom de MySQL no està oficialment exposada. Des de fa més de 10 anys, les eines han mantingut el prefix *My*. Es creu que té relació amb el nom de la filla del cofundador Monty Widenius, que es diu My.

Llenguatges de programació

Existeixen diverses APIs que permeten, a aplicacions escrites en diversos llenguatges de programació, accedir a les bases de dades MySQL, incloent C, C++, C#, Pascal, Delphi (via

[14] **GNU GLP:** és una llicència creada per Free Software Foundation a mitjans dels 80 i està orientada principalment a protegir la lliure distribució, modificació i ús de software.
[15] **ANSI C:** primera estandardització del llenguatge C
[16] **ISAM:** *Indexed Sequential Access Method* (Mètode d'accés Seqüencial Indexat), mètode per emmagatzemar informació.

dbExpress), Eiffel, Smalltalk, Java ... cada un d'aquest utilitza una API específica. També existeix una interfície ODBC, anomenada MyODBC que permet a qualsevol llenguatge de programació (PHP, java, C++, ASP, ...) que suporti ODBC comunicar-se amb les bases de dades MySQL.

Aplicacions

MySQL és molt utilitzat a aplicacions web com MediaWiki o Drupal, en plataformes (Linux/Windows-Apache-MySQLE-PHP/Perl/Python), i per eines de seguiment d'errors com Bugzilla. La seva popularitat com aplicació web està molt lligada a PHP, que sovint apareix en combinació amb MySQL. MySQL és una Base de Dades molt ràpida en la lectura quan utilitza el motor no transaccional MyISAM, però pot provocar problemes d'integritat a entorns d'alta concurrència en la modificació. A aplicacions web hi ha baixa concurrència en la modificació de dades i en canvi l'entorn és intensiu en lectura de dades, el que fa a MySQL ideal per aquest tipus d'aplicacions.

Especificacions

- *Plataformes:* AIX, BSD, FreeBSD, HP-UX, GNU/Linux, Mac OS X, NetBSD, Novell Netware, OpenBSD, OS/2 Warp, QNX, SGI IRIX, Solaris, SunOs, SCO OpenServer, SCO UnixWare, Tru64, Windows 95, Windows 98, Windows NT, Windows 2000, Windows XP, Windows Vista i altres versions de Windows.
- *Característiques de la versió 5.0.22:* té un ampli subconjunt d'ANSI SQL 99, i varies extensions, suport a multiplataforma, procediments emmagatzemats, triggers, cursors, vistes actualitzables, suport a VARCHAR, INFORMATION_SCHEMA, Mode Strict, Suport X/Open XA de transaccions distribuïdes (transacció en dos fases com part d'això) utilitzant el motor InnoDB d'Oracle, motors d'emmagatzemament independents (MyISAM per lectures ràpides, InnoDB per transaccions i integritat referencial), transaccions amb els motors d'emmagatzemament InnoDB, BDB i Cluster (punts de recuperació amb InnoDB), suport per SSL, query caching, sub-selects (o selects aniuats), llibreria de la Base de Dades *embedded*[17], suport complet per Unicode, conforme a les regles ACID utilitzant els motors InnoDB, BDB i Cluster,
- *Característiques addicionals:* Utilitza GNU Automake, Autoconf, i Libtool per a portabilitat, ús de multifils mitjançant fils de *kernel*, empra taules en disc b-tree per cerques ràpides amb compressió d'índex, taules *hash* a memòria temporals, el codi MySQL es proba amb Purify (un detector de memòria perduda comercial) així com amb Valgrind, una eina GPL, complet suport per operadors i funcions a clàusules *select* i *where*, complet suport per clàusules *group by* i *order by*, suport de funcions d'agrupació, en seguretat ofereix un sistema de contrasenyes i privilegis segur mitjançant verificació basada en el *host* i el tràfic de contrasenyes està xifrat al connectar-se a un servidor, suporta gran quantitat de dades (MySQL Server té bases de dades de fins 50 milions de registres), es permeten fins 64 índex per taula (32 abans de MySQL 4.1.2), cada índex pot consistir des d'1 fins 16 columnes o pars de columnes, el màxim ample de límit són 1000 bytes (500 abans de la versió 4.1.2), els clients es connecten al servidor utilitzant *sockets* TCP/IP en qualsevol plataforma, en sistemes Windows es poden connectar utilitzant *named pipes* i en sistemes Unix utilitzant fitxers *socket Unix*, a MySQL 5.0 els clients i servidors Windows es poden connectar utilitzant memòria compartida, MySQL conté el seu propi paquet de probes de rendiment proporcionat amb el codi font de la distribució de MySQL.

MySQL és un sistema d'administració de Base de Dades. Una Base de Dades és una col·lecció estructurada de taules que contenen dades. Aquesta pot ser des d'una simple llista de compres a una galeria de pintures o el vast volum d'informació a una xarxa corporativa. Per agregar, accedir i processar dades guardades a un computador, si es necessita un administrador com MySQL Sever.

[17] Embedded: "empotrades", "encapsulades".

Donat que els computadors són molt bons manegant grans quantitats d'informació, els administradors de bases de dades juguen un paper central en computació, com aplicacions independents o com parts d'altres aplicacions.

MySQL és un sistema d'administració relacional de Base de Dades. Una Base de Dades relacional arxiva dades a taules separades en comptes de col·locar les dades en un gran arxiu. Això permet velocitat i flexibilitat. Les taules estan connectades per relacions definides que fan possible combinar dades de diferents taules sobre comanda.

MySQL és software de font oberta. Font oberta significa que és possible per qualsevol persona utilitzar-ho i modificar-ho. Qualsevol persona pot descarregar el codi font de MySQL i utilitzar-ho sense pagar. Qualsevol interessat pot estudiar el codi font i ajustar-ho a les seves necessitats. MySQL s'utilitza el GPL (GNU General Public License) per definir que poden fer i que no poden fer am el software en diferents situacions. Si no s'ajusta al GLP o requereix introduir codi MySQL a aplicacions comercials, es pot comprar una versió comercial llicenciada.

Quina llicència utilitzar? La llicència GNU GLP de MySQL obliga a distribuir qualsevol producte derivat (aplicació) sota la mateixa llicència. Si un desenvolupador desitja incorporar MySQL al seu projecte però no desitja distribuir-lo sota llicència GNU GLP, pot adquirir una llicència comercial de MySQL que permet fer justament això.

En l'actualitat, la sèrie en desenvolupament de MySQL, és la 5.1 a la qual s'afegeixen noves característiques en relació a la sèrie 5.0. La sèrie de producció actual de MySQL és 5.0, la penúltima versió estable de la qual és la 5.0.26, llançada a octubre de 2006. Actualment, es pot descarregar la sèrie 5.0.27. La sèrie de producció anterior va ser la 4.1, versió estable de la qual és la 4.1.7 llançada a octubre de 2004. A aquestes versions de producció només s'arreglen problemes, és a dir, ja no s'afegeixen noves característiques. I a les versions anteriors només es corregeixen *bugs*[18] crítics.

3.3.2 PHP

És un llenguatge de programació utilitzat freqüentment per la creació de contingut per llocs web amb els quals es poden programar les pàgines html i els codis de font [8]. PHP és un acrònim recursiu que significa "PHP Hypertext Pre-processor" (inicialment PHP Tools o Personal Home Page Tools), i es tracta d'un llenguatge interpretat utilitzat per la creació d'aplicacions per a servidors, o creació de contingut dinàmic per a llocs web. Últimament també per a la creació d'altres tipus de programes incloent aplicacions amb interfície gràfica utilitzant les llibreries Qt o GTK+.

Una visió general

El fàcil ús i la similitud amb els llenguatges més comuns de programació estructurada, com C i Perl, permeten a la majoria dels programadors experimentats crear aplicacions complexes amb una corba d'aprenentatge molt suau. També es permet que es puguin involucrar amb aplicacions de contingut dinàmic sense haver d'aprendre tot un nou grup de funcions i pràctiques.

A causa del disseny de PHP, també és possible crear aplicacions amb una interfície gràfica per l'usuari (també anomenada GUI), utilitzant l'extensió PHP-Qt o PHP-GTK. També pot ser utilitzat des de la línia d'ordres, de la mateixa manera que Perl o Python poden fer-ho. Aquesta versió de PHP s'anomena PHP CLI (*Common Line Interface*).

La seva interpretació i execució es dóna en el servidor web, en el qual es troba emmagatzema l'script, i el client només rep el resultat de l'execució. Quan el client fa una petició al servidor per a que li enviï una pàgina web, generada per un script PHP, el servidor executa el intèrpret de PHP, el qual processa l'script sol·licitat que generarà el contingut de manera dinàmica, podent modificar el contingut a enviar, i retorna el resultat al servidor, el qual s'encarrega de retornar-ho al client. A

[18] Bugs: és un defecte del software, és el resultat d'un error o deficiència durant el procés de creació de programes d'ordinador.

més, és possible utilitzar PHP per generar arxius PDF, Flash, així com imatges en diferents formats, entre d'altres coses.

Permet la connexió a diferents tipus de servidors de bases de dades tals com MySQL, Postgres, Oracle, ODBC, DB2, Microsoft SQL Server, FireBird i SQLite; lo qual permet la creació d'aplicacions web molt robustes.

PHP també té la capacitat de ser executat en la majoria dels sistemes operatius tals com UNIX (y d'aquest tipus, com Linux o Mac OS X) i Windows, i pot interactuar amb els servidors de web més populars ja que existeix en versió CGI, mòdul per a Apache, i ISAPI.

El model PHP pot ser vist com una alternativa al sistema de Microsoft que utilitza ASP.NET/C#/VB.NET, a ColdFusion de la companyia Adobe (abans Macromedia), a JSP/Java de Sun Microsystems, i al famós CGI/Perl. Encara que la seva creació i desenvolupament es dóna a l'àmbit dels sistemes lliures, sota la llicència GNU, existeix a més un IDE comercial anomenat Zend Optimizer. Recentment, CodeGear (la divisió de llenguatges de programació de Borland) ha tret al mercat un entorn integrat de programació per a PHP, anomenat Delphi for PHP.

La Història

PHP va ser originalment dissenyat en Perl, seguit per l'escriptura d'un grup de CGI binaris escrits en el llenguatge C pel programador danès-canadenc Rasmus Lerdorf al any 1994 per a mostrar el seu currículum vitae i guardar diverses dades, com la quantitat de tràfic que la seva pàgina web rebia. El 8 de juny de 1995 va ser publicat "Personal Hom Page Tools" després de que Lerdorf ho combinés am el seu propi *Form Interpreter* per a crear PHP/FI.

Evolució de les versions:

- *PHP 3.2.4.3:* Dos programadors israelites del Technion, Zeev Suraski i Andi Gutmans, van reescriure l'analitzador sintàctic (*parser*) al l'any 1997 i van crear la base del PHP3, canviant el nom del llenguatge a la forma actual. Immediatament van començar experimentacions públiques de PHP3 i va ser publicat oficialment al juny del 1998. Per a 1999, Suraski i Gutmans van reescriure el codi de PHP, produint el que avui es coneix com Zend Engine o motor Zend, un *portmanteau* dels noms d'ambdós, Zeev i Andi. També van fundar Zend Technologies a Ramat Gan, Israel.
- *PHP 4:* Al maig de 2000 va ser llançat sota el poder del motor Zend Engine 1.0. L'última versió de PHP4 disponible al febrer de 2007 és la 4.4.7. PHP.NET va anunciar el dia 13 de Juliol de 2007 que la versió 4 de PHP ha quedat discontinuada.
- *PHP5:* el 13 de Juliol de 2004 va ser llançat, utilitzant el motor Zend Engine II. La versió més recent de PHP és la 5.2.3, que inclou totes les avantatges que proveeix el nou Zend Engine II, com: Suport sòlid per a Programació Orientada a Objectes amb PHP Data Objects, millores de rendiment, millor suport per MySQL amb extensió completament reescrita, millor suport a XML, suport natiu per SQLite, suport integrat per a SOAP, iteradors de dades, exepcions d'errors...
- *PHP6:* està previst el llançament aviat, quan es llanci aquesta nova versió, quedaran només dues rames actives en desenvolupament (PHP5 i 6).

Principals usos del PHP

- Programació de pàgines web dinàmiques, habitualment en combinació amb el motor de Base de Dades MySQL, encara que compta amb suport natiu per a altres motors, incloent l'estàndard ODBC, el que amplía en gran mida les seves possibilitats de connexió.
- Programació en consola, a l'estil Perl o Shell scripting.
- Creació d'aplicacions gràfiques independents del navegador, per mitjà de la combinació de PHP i Qt/GTK+, el que permet desenvolupar aplicacions d'escriptori en els sistemes operatius en els que estigui suportat.

Avantatges que ofereix PHP

- És un llenguatge multiplataforma.
- Té capacitat de connexió amb la majoria dels sistemes gestors de Base de Dades que s'utilitzen a l'actualitat, destaca la seva connectivitat amb MySQL.
- Llegeix i manipula dades des de diverses fonts, incloent dades que poden ingressar els usuaris des de formularis HTML.
- Capacitat d'expandir el seu potencial utilitzant l'enorme quantitat de mòduls (anomenats ext's o extensions).
- Posseeix una amplia documentació a la seva pàgina oficial, entre la qual es destaca que totes les funcions del sistema estan explicades i exemplificades en un únic arxiu d'ajuda.
- Es lliure, pel que es presenta com una alternativa de fàcil accés per a tothom.
- Permet les tècniques de Programació Orientada a Objectes.
- Permet crear formularis per a la Web.
- Té una biblioteca nativa de funcions summament àmplia i inclosa.
- No requereix definició de tipus de variables ni maneig detallat del baix nivell.

A continuació es presentarà a la *Figura 1* un exemple de codi en PHP i s'explicaran les curiositats.

```
<html>
<head>
    <title>Ejemplo</title>
</head>
<body>
<?php
if (isset($_POST['muestra'])) {
     echo 'Hola, '.htmlentities($_POST['nombre'])
         .', tu comida favorita es:'. htmlentities($_POST['comida']);
} else {
?>
<form method="POST" action="?">
    ¿Cuál es tu nombre?
    <input type="text" name="nombre">
    ¿Cuál es tu comida favorita?
    <select name="comida">
        <option>Spaguetis</option>
        <option>Asado</option>
        <option>Pizza</option>
    </select>
    <input type="submit" name="muestra" value="Seguir">
</form>
<?php
}
?>
</body>
</html>
```

Figura 1: Exemple de codi PHP

Les variables enviades per un formulari utilitzant el mètode POST, són rebudes en el llenguatge dins de la matriu $_POST, lo qual facilita l'obtenció d'aquest tipus de dades. Aquest mateix mètode es utilitzatpel llenguatge per a totes les fonts d'informació a una aplicació web, tals com *cookies*[19] a la matriu $_COOKIES, variables de URL a $_GET(que en formularis es poden servir per guardar les dades), les variables de sessió utilitzant $_SESSION, i variables del servidor i del client per mitjà de la matriu $_SERVER.

El codi PHP està incrustat dins de l'HTML i interactua amb el mateix, el que permet dissenyar la pàgina Web en un editor comú d'HTML i afegir el codi dinàmic dins de les etiquetes <?php ?>.

El resultat mostra i oculta certes porcions del codi HTML en forma condicional.

[19] **Cookie:** és un fragment d'informació que s'emmagatzema al disc dur del visitant d'una pàgina web a través del seu navegador, a petició del servidor de la pàgina.

Es possible utilitzar funcions pròpies del llenguatge per a aplicacions Web com *htmlentities()*, que converteix els caràcters que tenen algun significat especial en el codi HTML o que es podrien desplegar erròniament en el navegador com accents o dièresis, en els seus equivalents en format HTML.

Per finalitzar s'adjunta un quadre (*Taula 1*) amb la evolució històrica del PHP:

Versión	Fecha	Cambios más importantes
PHP 1.0	8 de Junio de 1995	Oficialmente llamado "Herramientas personales de trabajo (PHP Tools)". Es el primer uso del nombre "PHP".
PHP Version 2 (PHP/FI)	16 de Abril de 1996	Considerado por el creador como la "más rapida y simple herramienta" para la creación de páginas webs dinámicas .
PHP 3.0	6 de Junio de 1998	Desarrollo movido de una persona a muchos desarrolladores. Zeev Suraski y Andi Gutmans reescriben la base para esta versión.
PHP 4.0	22 de Mayo de 2000	Se agregan avanzadas de dos etapas analizar/ejecutar la etiqueta-análisis sistema llamado entorno motor Zend.
PHP 4.1	10 de Diciembre de 2001	Introducidas las variables superglobals ($_GET, $_SESSION, etc.)
PHP 4.2	22 de Abril de 2002	Se deshabilitan register_globals por defecto
PHP 4.3	27 de Diciembre de 2002	Introducido la CLI, en adición a la CGI
PHP 4.4	11 de Julio de 2005	
PHP 5.0	13 de Julio de 2004	Motor Zend II con un nuevo modelo de objetos.
PHP 5.1	25 de Noviembre de 2005	**Descontinuado el 13 de Julio de 2007**
PHP 5.2	2 de Noviembre de 2006	Habilitado el filtro de extensiones por defecto
PHP 5.2.3	31 de Mayo 2007	

Taula 1: Evolució històrica del PHP

3.4 Obstacles legals

Quan es va començar a desenvolupar aquest projecte tenia clares aspiracions empresarials, ja que fa poc més d'un any, la llei encara tenia un buit legal respecte al que Internet i els seguiment de premsa es refereix. Això va provocar les aspiracions de moltes empreses que ja es dedicaven al seguiment de premsa a la ràdio, televisió i sobretot premsa escrita. El naixement de nous portals web, amb la informació més actual possible, oferint una competència a la ràdio i a la televisió que la premsa tradicional no pot oferir. D'aquesta manera les empreses que ja es dedicaven al press-clipping van veure la possibilitat d'expandir els seus horitzons a través de la xarxa, ja que aquesta oferia unes expectatives de comerç amb moltes possibilitats i sense molta necessitat d'inversió, per comparativa amb la premsa escrita que cal comprar-la.

Però totes aquestes expectatives i il·lusions d'expansió cap a nous camps es va truncar quan els diferents mitjans i les diferents societats protectores dels drets d'autor d'Espanya es van començar a queixar sobre l'abús que s'estava fent de les notícies penjades a la xarxa. Ja que les empreses de press-clipping podien accedir a la informació de les webs i adquirir les notícies d'una manera gratuïta i sense mesura alguna per part dels autors (o mitjans) que les penjaven.

Així doncs, aquesta discussió sobre l'ètica de les empreses de recaptació de notícies i els mitjans que les distribueixen es va elevar al Congrés del Diputats, el qual va determinar el promoure modificacions a la *Llei de Protecció Intel·lectual* (LPI) que va donar lloc a que les empreses de press-clipping també haguessin de pagar un cànon per el recull d'informació d'Internet. La LPI diu: "*Quan es realitzin recopilacions d'artícles periodístics que consisteixen bàsicament en la seva reproducció i esmentada activitat es realitzi* ***amb fins comercials****, l'autor que no es s'hagi oposat expressament tindrà dreta a percebre una remuneració equitativa.*".

D'aquesta manera, la situació que es preveia com un gran esdeveniment per les empreses de seguiment de mitjans de comunicació es va quedar en una mera temptativa. Moltes, han optat per subcontractar l'obtenció d'aquestes notícies on-line per tal de no haver-hi de pensar i les que han optat per explotar aquest camp han hagut de realitzar una inversió major de la que preveien per dur a terme la recaptació de les notícies.

Finalment, degut a la LPI, l'empresa a on es va començar a realitzar aquest PFC va optar per la primera opció esmentada anteriorment, la de subcontractar el seguiment on-line a empreses que ja estaven especialitzades en el tema i no tenir problemes al respecte amb les possibles evolucions de la llei. D'aquesta manera, el projecte va deixar de ser un futur producte comercial per passar a ser un projecte.

4 Metodologia

Respecte a la metodologia que es pot emprar pel desenvolupament d'un projecte hi ha dos corrents clares al mercat, la Mètrica (que ja va per la versió 3) i la batejada com Extreme Programming. Aquestes dues tendències seran comentades a continuació per poder prendre la decisió de quina és la forma de treball que s'ajusta millor per poder fer una bona realització del projecte.

4.1 MÈTRICA Versió 3

La MÈTRICA Versió 3, és un instrument útil per a la sistematització de les activitats que donen suport al cicle de vida del software i permeten portar a terme els següents objectius:

Proporcionar o definir Sistemes d'Informació que ajudin a aconseguir les fites de l'organització, que l'està emprant, mitjançant la definició d'un marc estratègic per al desenvolupament dels mateixos [8].

- Dotar a l'organització de productes software que satisfacin les necessitats dels usuaris donant una major importància a l'anàlisi de requisits.
- Millorar la productivitat dels departaments de Sistemes i Tecnologies de la Informació i les comunicacions, permetent una major capacitat d'adaptació als canvis i tenint en compte la reutilització de software en la mida del possible.
- Facilitar la comunicació i enteniment entre els diferents participants en la producció de software al llarg del cicle de vida del projecte, tenint en comte el seu paper i responsabilitat, així com les necessitats de tots i cadascun d'ells.
- Facilitar l'operació, manteniment i ús dels productes software obtinguts.

Aquesta versió de MÉTRICA contempla el desenvolupament de Sistemes d'Informació per a les diferents tecnologies que actualment estan convivint i els aspectes de gestió que asseguren que un projecte compleixi els seus objectius en termes de qualitat, cost i terminis.

El seu punt de partida és la versió anterior de MÉTRICA de la qual s'han conservat l'adaptabilitat, flexibilitat, i senzillesa, així com l'estructura d'activitats i tasques, si bé les fases i mòduls de MÉTRICA versió 2.1 ha donat pas a la divisió den Processos, més adequada a l'entrada-transoformació-sortida que es produeix a cada una de les divisions del cicle de vida d'un projecte. Per a cada tasca es detallen els participants que intervenen, els productes d'entrada i de sortida així com les tècniques i pràctiques a emprar per la seva obtenció.

A l'elaboració de MÉTRICA Versió 3 s'ha tingut en compte els mètodes de desenvolupament més estesos, així com els últims estàndards d'enginyeria del software i qualitat, a més de referències específiques en quant a seguretat i gestió de projectes, tal com especifica el Ministeri d'Administracions Públiques (MAP), a través del Consell Superior d'Informàtica (CSI). També s'ha tingut en compte l'experiència dels usuaris de les versions anteriors per a solucionar els problemes o deficiències detectats.

En una única estructura la metodologia MÉTRICA Versió 3 cobreix diferents tipus de desenvolupament: estructurat i orientat a objectes, facilitant a través d'interfícies la realització dels processos de recolzament o organitzatius: Gestió de Projectes, Gestió de Configuració, Assegurament de Qualitat i Seguretat. L'automatització de les activitats proposades a l'estructura de MÉTRICA Versió 3 és possible ja que les seves tècniques estan suportades per una amplia varietat d'eines d'ajuda al desenvolupament.

Així els processos de l'estructura principal de MÉTRICA Versió 3 són els següents:

- **Planificació de Sistemes d'Informació:**

  L'objectiu és proporcionar un marc estratègic de referència pels Sistemes d'Informació d'un determinat àmbit de l'organització. El resultat del Pla de Sistemes ha de, per tant,

orientar les actuacions en matèria de desenvolupament de Sistemes d'Informació amb l'objectiu bàsic de recolzar l'estratègia corporativa, elaborant una arquitectura d'informació i un pla de projectes informàtics per a donar suport als objectius estratègics. Per aquest motiu es necessari un procés com el de Planificació de Sistemes d'Informació.

- **Desenvolupament de Sistemes d'Informació:**

  Conté totes les activitats i tasques que s'han de portar a terme per a desenvolupar un sistema, des de cobrir l'anàlisi de requeriments fins la instal·lació del software. A més de les tasques relatives a l'anàlisi, inclou dos parts en el disseny de sistemes: arquitectònic (que defineix la relació entre cada un dels elements estructurals del sistema) i detallat (que defineix el sistmea a tots els nivells, interfície, dades, estructura, ...).

- **Manteniment de Sistemes d'Informació:**

  L'objectiu d'aquest procés és l'obtenció d'una nova versió d'un sistema d'informació desenvolupat amb MÉTRICA, a partir de les peticions de manteniment que els usuaris realitzen amb motiu d'un problema detectat al sistema o per la necessitat de una millora del mateix. Com a conseqüència d'això, només es consideraran en MÉTRICA Versió 3 els tipus de Manteniment Correctiu i Evolutiu. S'exclouen els tipus de Manteniment Adaptatiu Perfectiu, que abasten activitats tals com la migració i la retirada de software que precisarien el desenvolupament d'un tipus de metodologia específica per a resoldre el seu objectiu.

4.2 Extreme Programming

La programació extrema o *eXtreme Programming* (XP) és un enfocament de la enginyeria del software formulat per *Kent Beck*, autor del primer llibre sobre la matèria, *Extreme Programming Explained: Embrace Change* (1999) [9]. És la més destacada dels processos àgils de desenvolupament de software. A l'igual que aquest, la programació extrema es diferencia de les metodologies tradicionals principalment en que es posa més èmfasis en l'adaptabilitat que en la previsibilitat.

Els defensors de XP consideren que els canvis de requisits sobre la marxa són un aspecte natural, inevitable i inclús desitjable del desenvolupament de projectes. Creuen que ser capaç d'adaptar-se als canvis de requisits en qualsevol punt de la vida del projecte és una aproximació millor i més realista que intentar definir tots els requisits al començament del projecte i invertir esforços després en controlar els canvis en els requisits.

Es pot considerar la programació extrema com l'adopció de les millores metodologies de desenvolupament d'acord al que es pretén portar a terme amb el projecte, i aplicar-ho de manera dinàmica durant el cicle de vida del software.

Les característiques fonamentals del mètode són:

- Desenvolupament iteratiu i incremental: petites millores, unes després de les altres.
- Proves unitàries continues: freqüentment repetides i automatitzades, incloent proves de regressió. S'aconsella escriure el codi de la prova abans de la codificació.
- Programació en parelles: es recomana que les tasques de desenvolupament es portin a terme per dos persones en un mateix lloc. Es suposa que la major qualitat del codi escrit d'aquesta manera – el codi es revista i discutit mentre s'escriu – és més important que la possible pèrdua de productivitat immediata.
- Freqüentment interacció de l'equip de programació amb el client o usuari. Es recomana que un representant del client treballi juntament amb l'equip de desenvolupament.
- Correcció de tots els errors abans d'afegir noves funcionalitats. Fer entregues freqüents.

- Refactorització del codi, és a dir, reescriure algunes parts del codi per augmentar la seva llegibilitat i manteniment però sense modificar el seu comportament. Les proves han de garantir que a la refactorització no s'ha introduït cap error.
- Propietat del codi compartida: en comptes de dividir la responsabilitat en el desenvolupament de cada mòdul en grups de treball diferents, aquest mètode promou que tot el personal pugui corregir i estendre qualsevol part del projecte. Les freqüents proves de regressió garanteixen que els possibles errors siguin detectats.
- Simplicitat en el codi: és la millor manera de que les coses funcionin. Quan tot funcioni es podrà afegir funcionalitat si es necessari. La programació extrema aposta que és més senzill fer quelcom simple i tenir una mica de treball extra per canviar-ho si es necessari, que realitzar alguna cosa complicada i potser no utilitzar-ho mai.

La simplicitat i la comunicació són extraordinariament complementaries. Amb més comunicació resulta més fàcil identificar què s'ha o què no s'ha de fer. Mentre més simple és el sistema, menys tindrà que comunicar-se sobre aquest, el que porta a una comunicació més completa, especialment si es pot reduir l'equip de programadors.

4.3 Decisions sobre la mètrica a emprar

Degut al caràcter especial d'aquest PFC, s'ha optat per una metodologia de desenvolupament una mica especial. S'ha escollit seguir que la manera més òptima de treball és la que ofereix l'Extreme Programming, ja que està encarada a un desenvolupament àgil i dóna un dinamisme que es creu que s'ajustarà perfectament al desenvolupament del projecte.

Tanmateix, cal dir que per poder estructurar la documentació paral·lela al desenvolupament del projecte s'agafaran les avantatges que ofereix la MÈTRICA Versió 3. Això s'ha determinat així després de valorar que aquesta mètrica donarà a projecte una visió més formal i estructurada que ajudarà a l'hora d'especificar la evolució que vagi succeint al llarg del desenvolupament del PFC.

5 Planificació de sistemes d'informació

5.1 Definició i organització del PSI

5.1.1 Especificació d'Àmbit i Abast

Per poder arribar a tenir una visió clara del que el present projecte vol abastar, cal tenir en compte l'entorn al que està subjecte. Internet ofereix uns avantatges per els nous horitzons empresarials que fa que les empreses tecnològiques posin les seves mires en aquesta xarxa. Això fa que sorgeixin idees com la d'aquest PFC que vol aprofitar aquests avantatges.

Abans que per raons legals (en el punt 3.4 esmentades) deixés de tenir aplicacions empresarials, l'àmbit d'aquest projecte volia aprofitar-se de les oportunitats que ofereix aquest nou model de negoci. Així doncs recolzant-se en el immens poder de difusió d'Internet, es vol arribar a incloure tants portals de mitjans de comunicació com sigui possible.

Aquest PFC donarà la oportunitat als usuaris de configurar a un llistat quins són els mitjans des d'on vol obtenir els seguiments. D'aquesta manera, a mida que es vagin analitzant nous portals, s'anirà donant la oportunitat, previ avís, als usuaris de contractar nous seguiments per als nous mitjans.

Per una altra banda, l'usuari té l'opció de configurar els descriptors que desitgi (amb un límit preestablert i informat al client prèviament). Aquest descriptors seran emmagatzemats a una Base de Dades on hi haurà tota la col·lecció de tots els clients amb la relació de mitjans a on es s'han d'anar a cercar.

5.1.2 Definició del Pla de Treball

El mètode de treball seguit, tal com es pot intuir per la metodologia escollida, és un pla de treball iteratiu e incremental. Això ve donat per el tipus de projecte que s'ha escollit, ja que només pels objectius establerts es poden preveure canvis i millores, adaptacions i retocs que s'hauran d'anar fent fins que s'arribi a la versió òptima.

D'aquesta manera, primer hi ha hagut una fase d'aprenentatge al nou medi de treball. Aprendre sobre el món periodístic, sobre la tecnologia escollida (PHP) i per últim, sobre tot el tema de patronatge[20]. Posteriorment, es va decidir la forma de treball amb aquestes eines i les possibles fases a seguir, aquestes es presenten a continuació amb corresponent diagrama de Gantt per reflectir la temporalització que s'ha estimat per portar a terme el projecte.

1a Fase: *aprenentatge.*
- Món periodístic.
- Món tècnic.

2a Fase: *decisions respecte a l'entorn de treball.*
- Hardware.
- Software.

3a Fase: *anàlisi de l'estructura d'una web.*
- Estructura general.
- Localització d'enllaços.
- Localització de la notícia.

4a Fase: *indexació a la Base de Dades.*
- Filtració segons descriptors.
- Emmagatzemament de la notícia.

[20] **Patronatge:** conjunt de patrons, s'està fent referència a als patrons que s'hauran d'elaborar per poder cercar dins del contingut de les diferents pàgines web.

5a Fase: *observació d'altres webs.*

Estructura general.

Comparativa amb la web inicial.

Decisions al respecte:

- Calen canvis a nivell de programació o Base de Dades?

6a Fase: *modificacions a nivell de Base de Dades i programació.*

Iteracions sobre les fases 3, 4 i 5.

A continuació es mostrarà l'esquema corresponent a aquestes fases especificades mitjançant el diagrama de Gantt següent.

	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
1	1a Fase: aprenentatge	23 días	lun 02/01/06	mié 01/02/06	
2	Món periodístic	7 días	lun 02/01/06	mar 10/01/06	
3	Notícies	1 día?	lun 02/01/06	lun 02/01/06	
4	RSS	2 días	mar 03/01/06	mié 04/01/06	3
5	Press-Clipping	4 días	jue 05/01/06	mar 10/01/06	4
6	Món tècnic	16 días	mié 11/01/06	mié 01/02/06	2
7	Llibreria cUrl	10 días	mié 11/01/06	mar 24/01/06	5
8	Expressions Regulars	6 días	mié 25/01/06	mié 01/02/06	7
9	2a Fase: decisions respec	7 días	jue 02/02/06	vie 10/02/06	1
10	Hardware	2 días	jue 02/02/06	vie 03/02/06	6
11	Software	5 días	lun 06/02/06	vie 10/02/06	10
12	3a Fase: anàlisi de l'estruc	15 días	lun 13/02/06	vie 03/03/06	9
13	Estructura general	5 días	lun 13/02/06	vie 17/02/06	11
14	Localització d'enllaços	5 días	lun 20/02/06	vie 24/02/06	13
15	Localització de la notícia	5 días	lun 27/02/06	vie 03/03/06	14
16	4a Fase: indexació a la Ba	30 días	lun 06/03/06	vie 14/04/06	12
17	Filtració segons descriptor	20 días	lun 06/03/06	vie 31/03/06	15
18	Emmagatzemament de la n	10 días	lun 03/04/06	vie 14/04/06	17
19	5a Fase: observació d'altre	30 días	lun 17/04/06	vie 26/05/06	16
20	Estructura general.	5 días	lun 17/04/06	vie 21/04/06	18
21	Comparativa amb la inicial	10 días	lun 24/04/06	vie 05/05/06	20
22	Decisions al respecte	15 días	lun 08/05/06	vie 26/05/06	21
23	6a Fase: modificacions a r	100 días	lun 29/05/06	vie 13/10/06	19
24	Iteracions sobre les fases	100 días	lun 29/05/06	vie 13/10/06	22

Taula 2: Fases i Subfases del Pla de Treball

Després d'observar la *Taula 2* a la que queden reflectides les fases del projecte, es mostren una sèrie de figures que contenent el diagrama de Gantt a on es pot veure la seqüència temporal des de l'inici al final de les fases i subfases que formen el Pla de Treball del projecte.

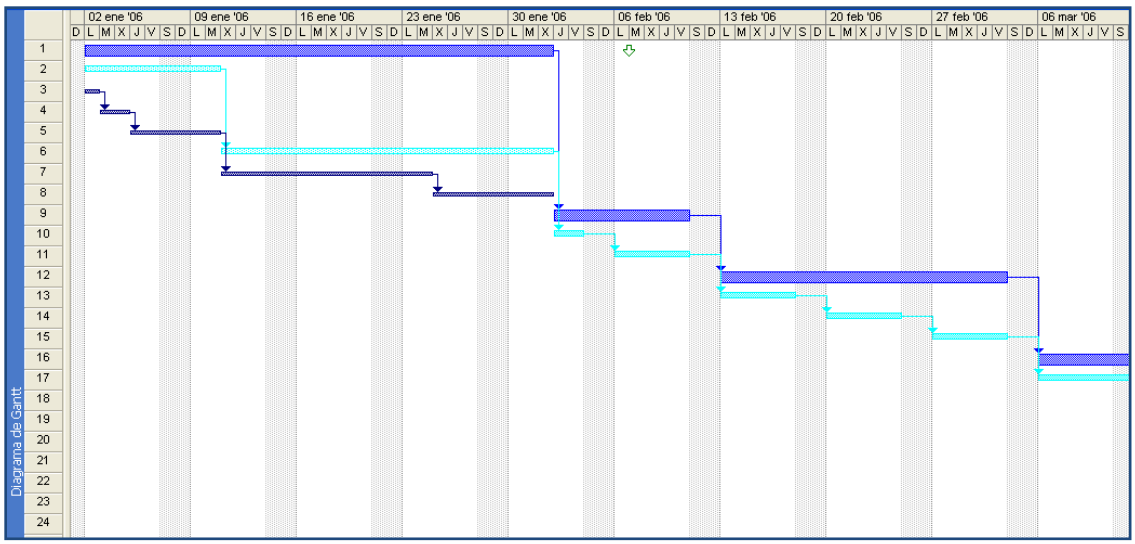

Figura 2: Fases 1, 2 i 3 del Pla de Treball

A la primera figura (*Figura 2*) es poden observar completament les 3 primeres fases, amb els seus corresponents apartats a portar a terme.

- Aprenentatge
- Decisions respecte l'entorn de treball
- Anàlisi de l'estructura d'una pàgina web.

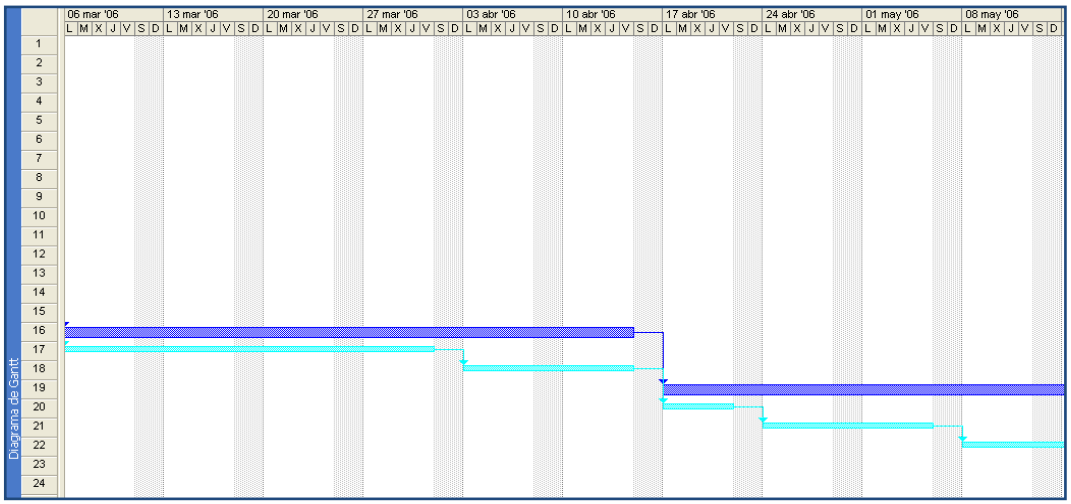

Figura 3: Fases 4 i 5 del Pla de Treball

A la *Figura 3* es pot veure la fase d'indexació i el principi de la fase d'observació de webs: Filtració i emmagatzematge i Estructura general, comparativa i decisions al respecte.

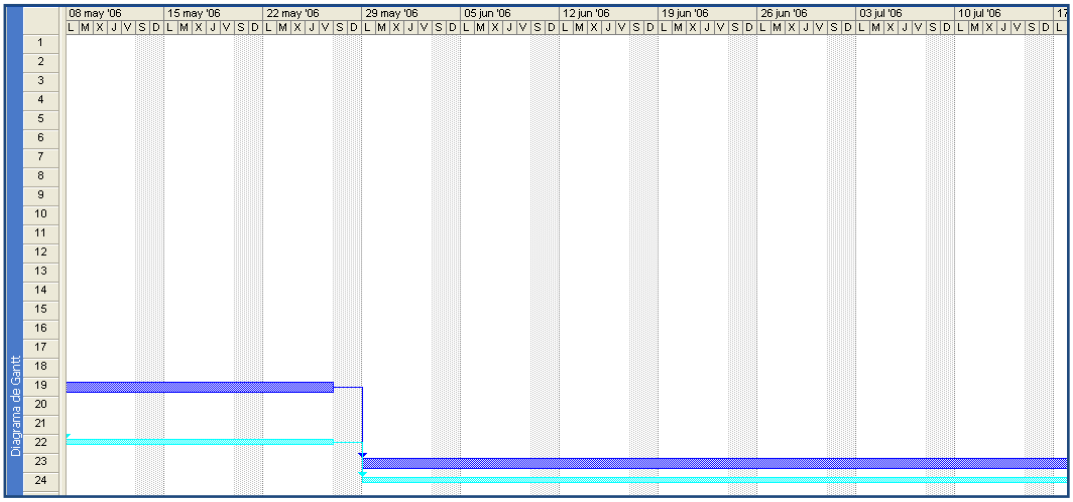

Figura 4: Fases 5 i 6 del Pla de Treball

A la *Figura 4* es pot observar el final de la fase 5 i el començament de la fase 6, a on es pot veure fins on dura la fase d'observació d'altres webs, la seva estructura general i la comparació amb la web primera per poder extreure les conclusions necessàries per poder fer els retocs necessaris a l'algorítmica de l'aplicació.

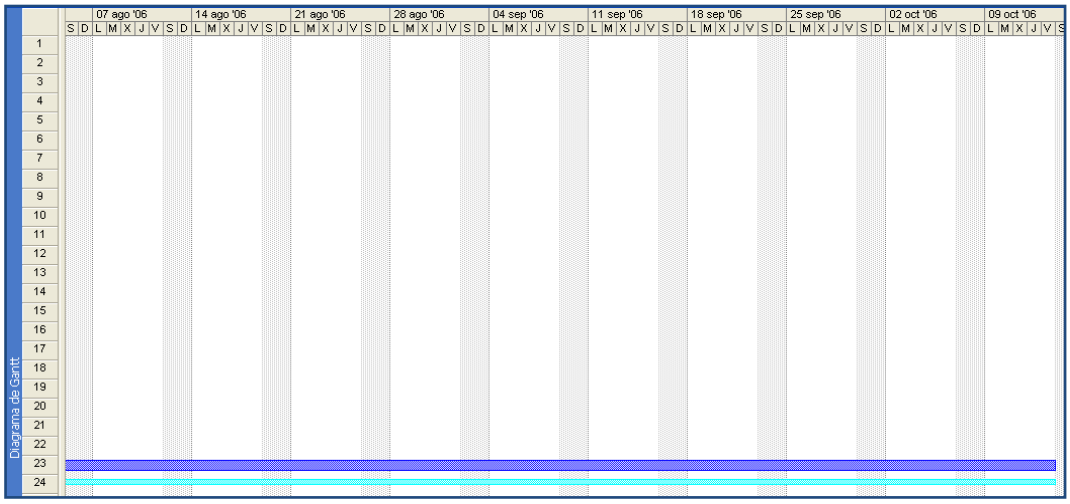

Figura 5: Última etapa de la Fase 6

Per últim, es pot observar a la *Figura 5* on es veu el final de la Fase 6 on es realitzaran les modificacions que pertoquin a la Base de Dades i a la programació del sistema, per poder arribar a una aplicació final que englobi tots els objectius esmentats al punt 2.2. d'aquesta memòria.

6 Viabilitat del Sistema

6.1 Establiment de l'Abast del Sistema

6.1.1 Estudi de la Sol·licitud

Durant el transcurs d'aquest apartat es volen especificar els requisits que són necessaris pel futur usuari del sistema. Per poder fer això, primerament es determinen els objectius a curt termini, a on cal tenir en compte les possibles ampliacions que s'aniran introduint a mida que vagi progressant i evolucionant el projecte.

Fins al moment, si es dóna un cop d'ull a les ofertes del mercat al voltant del tema del *press-clipping* es pot observar que existeixen empreses que faciliten seguiments de premsa, sobretot pel que fa a premsa escrita. El problema sorgeix quan es vol buscar un servei semblant que doni la possibilitat d'obtenir seguiments de notícies aparegudes a Internet, llavors la cerca es torna més complicada. Per això, la idea d'aquest projecte volia cobrir aquesta necessitat creant un nou portal.

Es vol aconseguir un portal on el client pugui tenir classificades les diferents famílies de notícies i pugui administrar els seguiments que sorgeixin. De tal manera, quan accedeixi a l'àrea de client es trobi amb la possibilitat d'escollir quins mitjans de comunicació vol que es recopilin notícies, també es trobarà amb l'opció d'administrar quins són els descriptors que té actius i possibilitat de gestionar-los com es vulgui.

6.1.2 Identificació de l'Abast del Sistema

Un cop establerta la sol·licitud es continua per la identificació de l'abast del sistema que es vol desenvolupar. Les accions que es poden portar a terme i que s'hauran de desenvolupar han de estar ben clares, per això s'aportarà a continuació una explicació detallada per millorar l'enteniment global del sistema i el seu abast.

Aquest projecte està dividit en dues àrees molt diferenciades. Aquestes es podrien catalogar com la interfície i l'algorítmica, ja que una és l'encarregada de mostrar al client els resultats de la segona.

El primer que es veu és la zona de treball del client. Aquesta, està formada per un portal des d'on l'usuari pot realitzar totes les seves consultes. Aquesta web està dividida en diferents zones:

- Les informatives:
  - Novetats
  - Serveis
  - Pàgina d'inici
- Les de treball pròpiament dites, on l'usuari pot gestionar:
  - Els seus descriptors
  - Els mitjans de comunicació als que està subscrits
  - Una bústia de suggeriments
  - L'àrea de visualització de les notícies recopilades per a ell.

En segon lloc es troba, l'anteriorment mencionada, àrea algorítmica. Aquesta porta a terme totes les tasques de recopilació. En sí, es la zona que treballa per la xarxa buscant informació i recopilant-la segons les necessitats de cada moment. Encara que es la part del projecte que treballa en *back-office*[21], és la major responsable de que el projecte es dui a terme, perquè sense el mòdul cercador que automatitzi tot el referent a la recopilació de la informació de la xara no existiria l'àrea d'investigació acadèmica que es vol aportar.

[21] **Back-office:** "darrera de l'usuari". Tasques que es realitzen sense que l'usuari sigui conscient de que se estan portant a terme

Tot seguit s'observa la *Figura 6* on es pot veure reflectit l'abast del que vol aportar i com està estructurat.

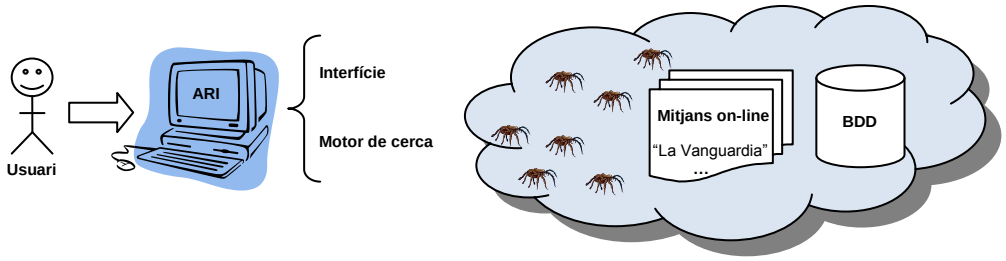

Figura 6: Estructura de l'Abast del Sistema

6.2 Estudi de la Situació Actual

6.2.1 Valoració de l'Estudi de la Situació Actual

Com a conclusió sobre la situació actual, es pot dir que com a idea inicial de projecte empresarial va ser una idea molt bona. Però, al sorgir inconvenients de caire legals, es va haver d'assumir un canvi de visió que ha fet que el projecte passés a ser d'investigació i estrictament acadèmic.

Fins al moment, les empreses no tenen una visió clara sobre les expectatives que ofereix Internet i per això hi ha un buit d'oferta a l'usuari en el que respecta al tema d'aquest projecte. A Espanya existeixen poques empreses que es dediquin a l'explotació d'aquest sector, com iMente o Anpro21[22] ofereixen dintre del seu catàleg de productes solucions semblants a la que aquest projecte pretén donar sortida.

6.2.2 Identificació dels Usuaris participants en l'Estudi de la Situació Actual

Per poder arribar a bon port s'ha investigat prèviament i s'ha analitzat les necessitats i els recursos de tots els usuaris participants de la situació actual. Com va haver-hi un abans i un després del sorgiment de la Llei de Protecció Intel·lectual (LPI), hi ha usuaris que van participar als inicis que després es van desentendre del projecte. De tota manera, s'ha decidit esmentar-los.

En un inici, va haver-hi el responsable del projecte, el cap de l'empresa que el desenvoluparia. Aquest usuari va ser qui va determinar els protocols i terminis per a portar a terme el projecte, així doncs, el quadre vist al punt 5.1.2 (Pla de Treball) va ser establert per aquest usuari i respectat posteriorment, encara que no es va arribar a desenvolupar tot el projecte sota les seves ordres.

Per un altre costat està la persona que desenvolupa el projecte (jo), que en un principi va començar a realitzar les tasques a fer a les ordres del cap de l'empresa, seguint les seves directrius i quan la LPI va canviar i la viabilitat del projecte a nivell empresarial es va veure afectada, vaig adoptar les responsabilitats del desenvolupament del projecte i va passar a ser de caire acadèmic.

6.2.3 Descripció dels Sistemes d'Informació Existents

Tal com s'ha comentat prèviament, no hi ha gaires empreses a Espanya que realitzin seguiments de premsa orientats a donar un servei específic a l'usuari i els que existeixen només se'ls poden permetre empreses amb una capacitat econòmica important i no estan encarats pels usuaris comuns.

Així es pot observar com el procediment actual que té l'usuari normal per poder satisfer la seva curiositat sobre un determinat tema es reflexa en el següent esquema.

[22] **iMente:** http://www.imente.com/
anpro21: http://www.anpro21.com/

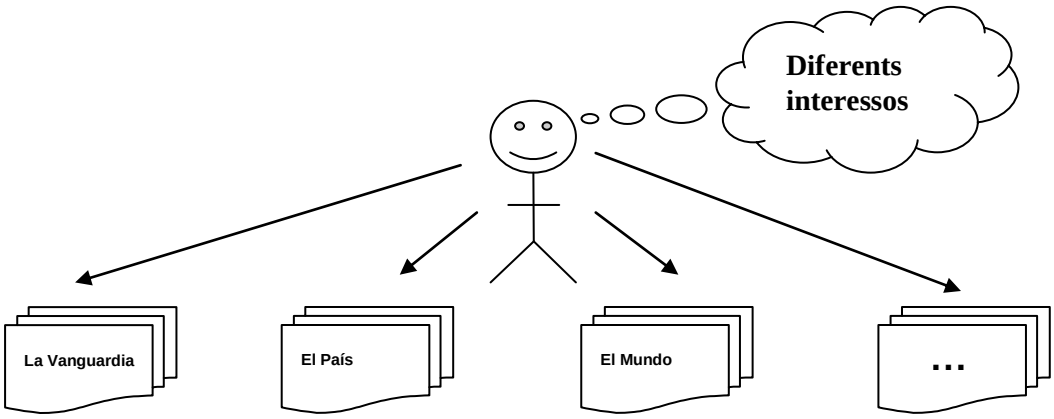

Figura 7: Diagrama de context del sistema actual existent

Aquest diagrama de context (*Figura 7*) vol explicar com fins al moment, l'usuari cerca per la xarxa en busca de la informació que l'interessa. D'aquesta manera, ha d'anar manualment pàgina per pàgina de cada portal web del mitjà de comunicació que li interessi i decidir si la notícia que està llegint és el que està buscant o no.

A continuació es reflexa aquest procediment a un nivell físic:

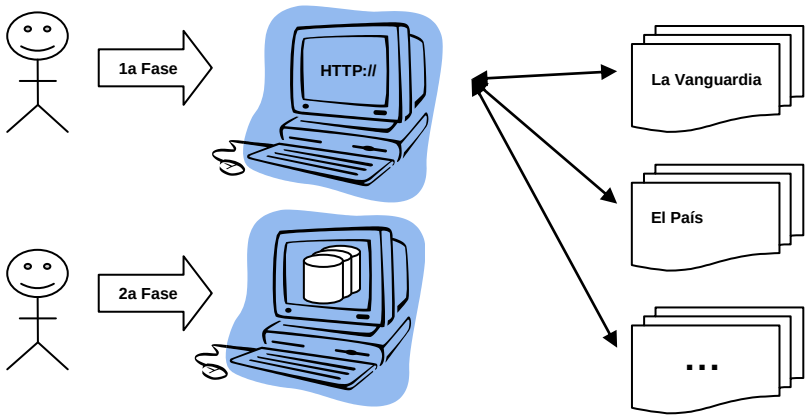

Figura 8: Model físic del sistema actual

Com es pot observar a la *Figura 8*, el procés actual de recopilació que ha de portar a terme l'usuari està dividit en dues fases:

- **Primera Fase:** visitar les webs dels portals de comunicació dels que es vol obtenir la informació per seleccionar les notícies que es volen sobre el tema que s'escull.
- **Segona Fase:** recopilar i emmagatzemar les notícies escollides d'alguna manera al seu ordinador personal.

6.3 Definició de Requisits del Sistema

Per poder fer una bona definició dels requisits del sistema, cal que fer un anàlisi exhaustiu de les necessitats de l'usuari de l'aplicació. Aquest projecte es realitza amb la voluntat de alliberar-lo del tràfec de cercar per les diferents pàgines web dels mitjans de comunicació, havent d'entrar a cada secció de cada mitjà investigant les notícies per extreure la informació desitjada.

Com a part dels requisits que sorgeixen al planificar aquest projecte es troba que el portal que es dissenya pensant en oferir ajut a l'usuari té dos seccions molt diferenciades: la part pública i la part privada.

6.3.1 Definició dels requisits de la Part Pública del Sistema

Aquesta àrea del projecte està encarada a facilitar a l'usuari la informació bàsica necessària per tenir una visió general de que és el projecte i que el pot oferir a nivell de prestacions. Per poder aportar tot això, es divideix la Part Pública en un seguit de seccions per fer més usable i amigable la seva navegació. A continuació es presenten les seccions i els seus requisits.

Secció 1: Pàgina Principal

A aquesta secció és la primera impressió que s'emporten els usuaris quan entren. Per aquesta raó els requisits d'aquesta secció estan molt clars: oferir una visió clara i concisa del que pot aportar aquest PFC a l'usuari si decideix utilitzar-ho. Oferir una navegació intuïtiva per tota la Part Pública, aportar notícies sobre la situació del Sistema i dades d'interès.

Secció 2: Novetats

Aquí es localitza tota la informació sobre els avenços que va adquirint el sistema. Ja sigui a l'àrea del sistema de serveis, com de mitjans de comunicació oferts. Els requisits que impliquen aquestes necessitats d'informació s'han de reflectir a la secció d'una manera clara i concisa. També cal que, tal com succeeix a la Pàgina Principal, es pugui navegar d'una manera senzilla cap a les altres zones de la Part Pública.

Secció 3: Serveis

La secció que aquí s'observa aportarà a l'usuari un catàleg de possibilitats de serveis que dóna l'aplicació que es trobarà, un cop s'hagi donat d'alta al sistema. De la mateixa manera que a les dos seccions anteriors, aquesta, ha de donar la possibilitat d'una navegació per totes les diferents seccions d'aquesta part.

Secció 4: Alta de nous usuaris

Aquesta àrea pública és la responsable de donar la possibilitat a l'usuari de donar-se d'alta al Sistema. Els requisits de la secció són els següents:

- Un formulari amb les dades que necessita el sistema per emmagatzemar:
  - Nom i Cognoms
  - Direcció
  - Telèfon i Fax
  - Correu electrònic
  - Nom d'usuari i contrasenya
  - Si desitja o no rebre correus electrònics informatius
- Enllaços cap a les diferents seccions per facilitar la navegabilitat.

Secció 5: Identificació dels usuaris

La secció que aquí s'especifica apareix en dos ocasions al llarg de la Part Pública. Primerament es troba formant part de la Pàgina Principal, per oferir a l'usuari un accés directe a la Part Privada. També, des de les diferents seccions, com Novetats o Serveis, es pot trobar un enllaç cap a aquesta secció que ofereix a l'usuari la possibilitat d'accedir a la seva àrea personal del portal web.

6.3.2 Definició dels requisits de la Part Privada del Sistema

Els requisits que es troben a aquesta part del portal que es desenvolupa pel projecte venen definits per les funcionalitats bàsiques que es volen oferir a l'usuari. Aquestes són sobretot les de que puguin accedir als seus seguiments i les notícies que aquests tenen. També un requisit important d'aquesta Part Privada és l'oferir al client una zona de gestió tant dels mitjans de comunicació als que s'està subscrit com dels descriptors dels quals vol fer els diferents seguiments.

Així doncs, les seccions que es requereixen a aquesta Part Privada són les que a continuació s'especifiquen.

Secció 1: Informació general

Només d'entrar a la Part Privada del projecte, el requeriment bàsic és la informació que hi trobarà l'usuari. Així doncs s'hauria de poder observar totes les dades bàsiques que es trobaran a les demés seccions, així com la informació més bàsica de la Part Pública que es pugui requerir.

Secció 2: Gestió de Descriptors

A aquesta secció l'usuari ha de poder gestionar els ítems (o descriptors) dels quals vol fer el seguiments i per tant es necessària que aquesta àrea de la Part Privada deixi a l'usuari donar d'alta i eliminar els descriptors que vulgui incorporar al sistema o eliminar-ho del mateix.

Secció 3: Gestió de Mitjans de Comunicació

A la secció de Gestió de Mitjans de Comunicació es necessari que l'usuari pugui realitzar les mateixes accions que amb els descriptors. Cal oferir al client del projecte un gestor dels mitjans que té a la seva disposició, informació dels nous mitjans i les possibilitats que tenen.

Secció 4: Gestió de Notícies

La secció de Gestió de Notícies ha d'oferir a l'usuari un llistat de totes les notícies que s'han recopilat per a ell i la opció d'administrar-les d'alguna manera. També cal que es pugui visualitzar completament el detall de les mateixes.

Secció 5: Bústia de Suggeriments

Com a última secció que cal oferir a l'usuari d'una secció que el permeti comunicar-se amb el coordinador del projecte, és a dir una mateixa. Aquí es podrà trobar un formulari que farà de pont entre els responsables de la gestió del projecte i els usuaris del portal.

6.4 Estudi d'Alternatives de Solució

6.4.1 Preselecció

Un cop s'ha especificat els requisits que són necessaris pel client, i abans d'implementar definitivament el projecte que es presenta a aquesta memòria, es fa necessari un estudi de totes les possibles solucions que poden existir al mercat pels requeriments esmentats al punt anterior.

Les alternatives trobades passen per la contractació de serveis com els que ofereix iMente, que porten endavant solucions semblants a les que ofereix aquest projecte, però d'unes dimensions bastant més grosses i ambicioses. Al ser una empresa consolidada dins el món del *press-clipping*, fa que hagi pogut assumir el increment econòmic que significava el canvi de la LPI.

6.4.2 Descripció, valoració i selecció

L'alternativa, per poder ser una opció a tenir en compte, hauria d'adaptar-se a les necessitats explicades al punt dels requisits del sistema. Necessitaria solucionar tots els problemes sobretot a nivell de senzillesa i transparència en la seva utilització, millorant el que el present projecte ofereix a l'usuari.

Segurament, iMente aporta la solució més eficaç i ràpida, però també amb uns costos adjunts. Al perdre el caire empresarial, aquest projecte vol investigar la problemàtica que hi ha darrere dels estàndards d'HTML que fins al moment és un terreny on no hi ha consens en el món de la programació.

La solució escollida ha estat la de continuar amb el desenvolupament del projecte, ja que el que es vol no és obtenir una gran solució empresarial, sinó, una investigació del món d'Internet i el codi HTML en el que es basa.

S'assumeix que la millor solució, la que ofereix el resultat final que es vol obtenir amb aquest projecte, la dóna iMente, (per un preu no assumible pel petit usuari, és una solució encarada a donar solucions a empreses). Però amb el desenvolupament d'aquest projecte es vol investigar quin nivell de dificultat té l'explotació de la informació a la Xarxa.

Primera Versió del desenvolupament

7 Anàlisi i Disseny del Sistema d'Informació (v.1)

7.1 Definició del Sistema (v.1)

7.1.1 Determinació de l'Abast del Sistema (v.1)

En aquest moment de l'estudi cal analitzar els requisits que s'han de solucionar i es volen desenvolupar. Per això cal determinar l'abast definitiu que es vol aconseguir pel sistema. És a dir, aquest sistema ha de donar la possibilitat d'obtenir un seguit de notícies que compleixin els desitjos establerts per l'usuari final. Aquests es poden resumir en:

- *Obtenir notícies*: aquest és el fi bàsic del projecte, el motiu pel qual es desenvolupa i es vol investigar la Xarxa, per esbrinar la dificultat que comporta el voler donar aquest servei a l'usuari.
- *Gestionar els mitjans de comunicació*: per poder portar una gestió i ordre de les notícies que obté cada usuari, cal facilitar-li la gestió dels mitjans de comunicació que té al seu abast per fer els seguiments de premsa que desitgi.
- *Gestionar els descriptors:* així com succeeix amb la gestió anterior, aquesta és una conseqüència lògica del primer punt, aquí esmentat, ja que per poder fer seguiments de premsa calen uns descriptors, que seran els ítems a localitzar dins de les notícies per poder filtrar-les com a bones per a l'usuari.

Aquestes tres àrees que s'ofereixen al client han de tenir associades uns processos específics. Pel cas de les dues gestions, aquest processos seran paral·lels a ambdós casos: donar d'alta un mitjà o descriptor, donar-lo de baixa, i la possibilitat de consultar-lo.

En canvi, l'obtenció de la notícia és el punt que més evolucionarà i que es podrà anar observant al llarg de les versions següents del desenvolupament. D'aquesta manera, a continuació (*Figura 9*), es presentarà el primer Diagrama de Context del Sistema a on es deixarà patent aquesta primera versió del projecte.

Diagrama de Context del Sistema

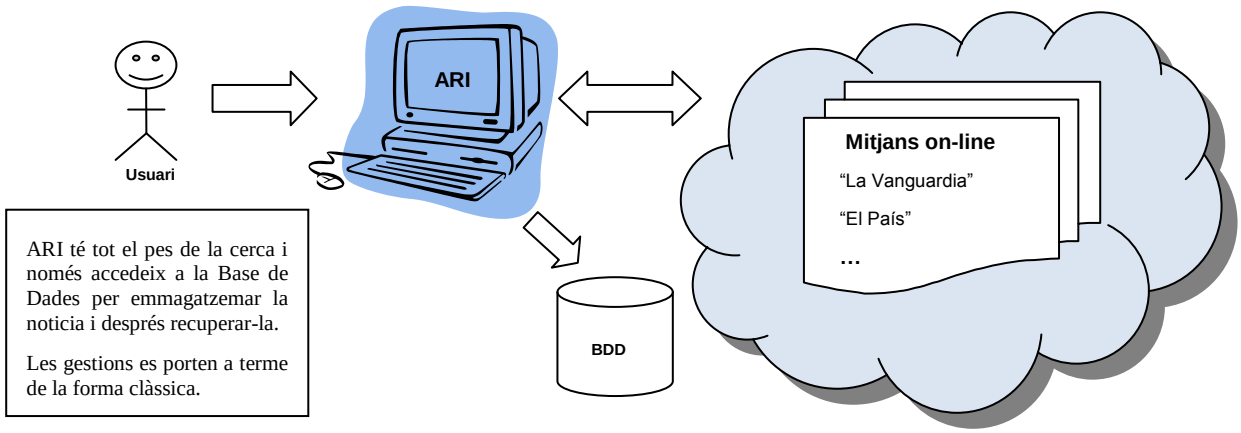

Figura 9: Diagrama de Context del Sistema (Versió 1)

7.1.2 Identificació d'Usuaris Participants i Finals (v.1)

Els usuaris participants en el procés de disseny del projecte en aquest punt del desenvolupament del sistema són principalment dos:

- El responsable de l'empresa que ha iniciat el projecte i ha determinat quines són les prioritats d'implementació i d'obtenció de mitjans de comunicació.
- Jo mateixa, com a desenvolupadora del projecte i co-responsable del mateix.

7.2 Establiment de Requisits (v.1)

Per poder fer un establiment definitiu dels requisits que ha de complir el projecte, s'ha de tenir una visió clara de quins són els possibles problemes als que s'haurà de fer front i que fins el moment, l'usuari afrontava manualment, és a dir, anava per Internet recorrent les diferents pàgines web dels mitjans de comunicació cercant les notícies que s'ajustessin als seus desitjos.

D'aquesta manera, l'objectiu bàsic a complir és trobar la forma de donar un servei a l'usuari tant senzill i efectiu com sigui possible. Per poder valorar la solució com òptima, l'usuari de l'aplicació ha de tenir la sensació que li és més rendible que el navegar ell mateix per les pàgines web dels mitjans de comunicació.

Així doncs, es troben dos classes d'obstacles a la cerca manual de les notícies que es vol optimitzar:

- La memòria, haver de recordar a cada moment quins mitjans de comunicació s'han visitat, quins paràmetres s'han investigat i quins falten, etc.
- L'espai, un cop es tenen les notícies, què es fa d'elles? Les ha de llegir totes i retenir que es diu, ha d'escriure i guardar els enllaços on les ha trobades o se les ha de descarregar al seu PC?

Els requisits que ha de complir el projecte han d'anar orientats a mitigar aquests dos problemes i fer que l'usuari es trobi amb un portal on aquest dos problemes exposats aquí es redueixin fins fer que l'obtenció de les notícies que desitja sigui un procés senzill i ergonòmic.

Per tant, els requisits generals que es troben són els següents:

- La transparència d'accés i gestió per a tenir organitzat tot el referent als mitjans de comunicació als quals s'està accedint durant el procés de recerca, el mateix al que ítems o descriptors es refereix, han d'estar introduïts a la Base de Dades.
- Una Base de Dades que recolzi tot el procés d'una manera eficaç i rendible, que faci de tot el procés de gestió quelcom natural. Pel que fa a l'obtenció de notícies, cal una estructura que afronti les dificultats de tenir administrat un gran volum d'informació, tota relacionada entre sí.
- Una algorísmica eficient, que faci que la cerca per la Xarxa sigui el més espontani i ràpid possible.

Un cop es tenen els requeriments bàsics del Sistema, es passa a fer una especificació dins de l'àmbit de cada procés que es desenvolupa a aquest projecte.

7.2.1 *La Gestió de Descriptors (v.1)*

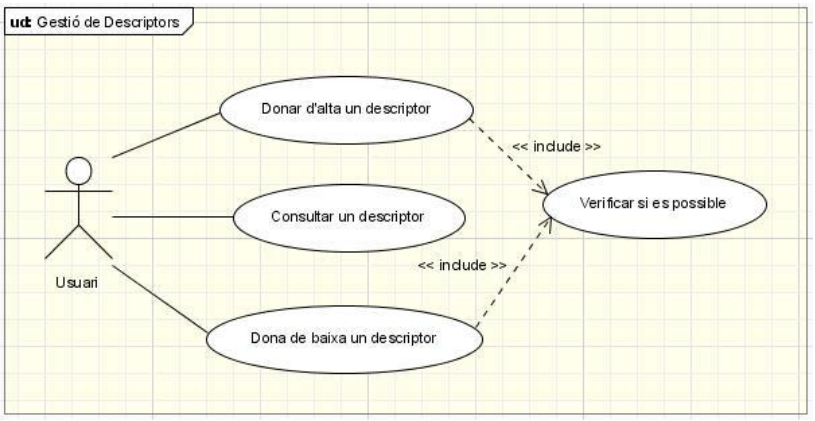

Figura 10: Cas d'ús – Gestió de Descriptors

A aquest diagrama, la *Figura 10* aquí representada, es pot veure en global els requisits funcionals que ha de facilitar la gestió de descriptors. En aquest cas, com es pot observar, només hi ha un tipus d'actor relacionat, l'usuari.

Fitxes dels casos d'ús associats al diagrama [10]

Cas d'ús	Alta d'un descriptor
Versió	1
Descripció	Donar d'alta un nou descriptor.
Actors	Usuari.
Precondició	No ha superat el límit de descriptors al sistema.
Flux Principal	1. Entrar a l'àrea del portal corresponent. 2. Introduir la informació del nou descriptor. 3. Donar d'alta el nou descriptor. a. Mirar si el descriptor ja està al sistema, si hi és associar-lo a l'usuari. b. Si no hi és al sistema, donar-ho d'alta al sistema i associar-ho a l'usuari.
Postcondició	Descriptor introduït al sistema.

Taula 3: Fitxa – Alta d'un descriptor

En aquesta primera versió de com es dóna d'alta un descriptor al sistema es poden observar els següents requisits:

- L'actor que realitza l'acció de donar d'alta un descriptor és l'usuari.
- Cal tenir en compte si el descriptor ja ha estat introduït al sistema per un altre usuari i actuar en conseqüència:
  - Si ja hi és, associar-ho a l'usuari que està realitzant la petició.
  - Si no hi és al sistema, introduir-ho i associar-ho a l'usuari.

Un cop es tenen en compte tots els requisits per donar d'alta un descriptor, aquest passarà a formar part del sistema i s'utilitzarà en la cerca i obtenció de notícies per la xarxa.

Cas d'ús	Consulta d'un descriptor
Versió	1
Descripció	Consultar un descriptor.
Actors	Usuari.
Precondició	Tenir algun descriptor introduït al sistema.
Flux Principal	1. Entrar a l'àrea del portal corresponent. 2. Observar les dades del descriptor.
Postcondició	Dades del descriptor consultades al sistema.

Taula 4: Fitxa – Consulta d'un descriptor

Cal que el cas d'ús de consulta d'un descriptor compleixi els següents requisits per l'execució:

- L'actor que realitza l'acció de consulta d'un descriptor és l'usuari.
- Cal tenir en compte que només es visualitzarà informació si l'usuari ha introduït prèviament algun descriptor al sistema.

Cas d'ús	Baixa d'un descriptor
Versió	1
Descripció	Donar de baixa un descriptor.
Actors	Usuari.
Precondició	El descriptor es troba prèviament introduït al sistema.
Flux Principal	1. Entrar a l'àrea del portal corresponent. 2. Escollir el descriptor a donar de baixa. 3. Donar de baixa el descriptor associat al l'usuari. a. Si hi ha més usuaris que l'utilitzen no esborrar del sistema. b. Si no hi ha més usuaris associats al descriptor, esborrar-ho del sistema.
Postcondició	Descriptor esborrat del sistema.

Taula 5: Fitxa – Baixa d'un descriptor

L'últim cas d'ús referent a la gestió de descriptors fa referència a donar de baixa un descriptor. Aquest cas ha de complir un seguit de requeriments:

- L'actor que realitza l'acció de donar de baixa un descriptor és l'usuari.
- Si el descriptor està associat a un altre usuari cal actuar en conseqüència:
  - Si està associat, esborrar l'enllaç entre el descriptor i l'usuari que dóna de baixa.
  - Si no hi està, esborrar l'enllaç i el descriptor del sistema.

D'aquesta manera, els requisits de la gestió de descriptors queden totalment analitzats.

7.2.2 La Gestió dels Mitjans de Comunicació (v.1)

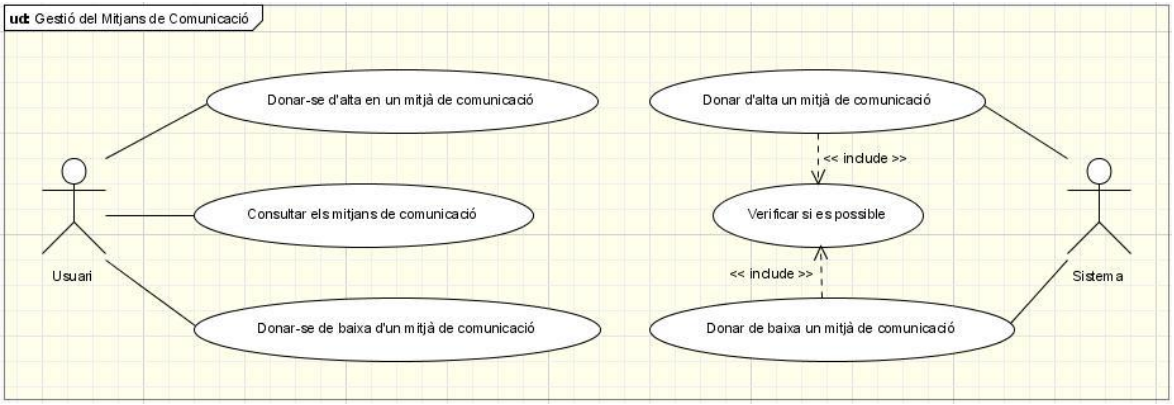

Figura 11: Cas d'ús – Gestió de Mitjans de Comunicació

A aquesta *Figura 11*, el diagrama que representa la gestió de mitjans de comunicació, on es pot veure que hi intervenen dos actors per portar a terme totes les tasques requerides.

Fitxes dels casos d'ús associats al diagrama

Cas d'ús	Alta d'un mitjà de comunicació
Versió	1
Descripció	Donar d'alta un mitjà de comunicació.
Actors	Sistema.
Precondició	El sistema està connectat a la xarxa.
Flux Principal	1. Descarregar el codi font de la pàgina principal del mitjà de comunicació 2. Analitzar el codi de la pàgina principal. a. Estructura dels enllaços. b. Confecció del patró per localitzar-los. 3. Descarregar el codi d'una notícia del mitjà de comunicació. a. Estructura d'una notícia. b. Confeccionar els patrons de les parts de la notícia. 4. Confeccionar el codi de manera que segueixi els patrons.
Postcondició	Codi confeccionat de manera que cerqui mitjançant els patrons.

Taula 6: Fitxa – Alta d'un mitjà de comunicació

Aquesta fitxa (*Taula 6*) fa referència als requisits de la gestió de mitjans de comunicació que es centren en els procediments que ha de seguir el sistema per poder obtenir un patró de cerca a un mitjà de comunicació. Aquests patrons són introduïts al sistema de cerca (que no a la Base de Dades) per poder realitzar la recaptació de la informació als portals desitjats.

Cas d'ús	Baixa d'un mitjà de comunicació
Versió	1
Descripció	Donar de baixa un mitjà de comunicació.
Actors	Administrador del Sistema.
Precondició	No hi ha.
Flux Principal	1. Eliminar els patrons del codi.
Postcondició	Codi confeccionat de manera que cerqui mitjançant els patrons.

Taula 7: Fitxa – Baixa d'un mitjà de comunicació

Referent a donar de baixa un mitjà de comunicació, en aquests moments l'administrador del sistema només ha d'eliminar els patrons corresponents del sistema per a que aquest no pugui reconèixer cap estructura i no recopili ni analitzi cap informació.

Cas d'ús	Alta en un mitjà de comunicació
Versió	1
Descripció	Donar-se d'alta en un mitjà de comunicació.
Actors	Usuari.
Precondició	El mitjà desitjat es troba prèviament recaptat pel sistema i l'usuari encara no s'ha donat d'alta.
Flux Principal	1. Entrar a l'àrea del portal corresponent. 2. Escollir del llistat quin mitjà de comunicació es desitja. 3. Donar d'alta al sistema l'associació entre el mitjà i l'usuari.
Postcondició	Usuari associat al mitjà de comunicació.

Taula 8: Fitxa – Baixa d'un mitjà de comunicació

A partir d'aquesta fitxa, l'actor que dur ha de dur a terme els requisits canvia, passa a ser l'usuari del sistema. Així doncs, l'usuari que es vol donar d'alta a un mitjà de comunicació ha d'anar a la zona reservada per la gestió de mitjans de comunicació i escollir del llistat el mitjà de comunicació del que desitja recaptar informació. El sistema emmagatzemarà la relació entre l'usuari i el mitjà de comunicació.

Cas d'ús	Consulta dels mitjans de comunicació
Versió	1
Descripció	Consulta d'un mitjà de comunicació.
Actors	Usuari.
Precondició	El mitjà de comunicació es troba prèviament a la llista de possibles mitjans.
Flux Principal	1. Entrar a l'àrea del portal corresponent. 2. Localitzar la taula de mitjans de comunicació. 3. Consultar les dades del mitjà que es desitja.
Postcondició	Mitjà de comunicació consultat.

Taula 9: Fitxa – Consulta dels mitjans de comunicació

La consulta d'un mitjà de comunicació és una acció que no necessitat gaires requisits, ja que l'usuari només ha d'entrar a la zona de gestió de mitjans i allà hi troba tot el referent als mitjans de comunicació que hi ha.

Cas d'ús	Baixa en un mitjà de comunicació
Versió	1
Descripció	Donar-se de baixa d'un mitjà de comunicació.
Actors	Usuari.
Precondició	El mitjà es troba associat a l'usuari dins del sistema.
Flux Principal	1. Entrar a l'àrea del portal corresponent. 2. Escollir el mitjà del que es vol donar de baixa. 3. Eliminar la relació entre el mitjà de comunicació i l'usuari del sistema.
Postcondició	Associació entre l'usuari i el mitjà esborrada del sistema.

Taula 10: Fitxa – Baixa en un mitjà de comunicació

Per últim, la fitxa que correspon a l'acció de donar-se de baixa, per part d'un usuari, d'un mitjà de comunicació. Aquesta necessita acció necessita que prèviament l'usuari hagi estat associat a un mitjà de comunicació per tal de poder eliminar quelcom del sistema.

7.2.3 L'obtenció de Notícies (v.1)

Figura 12: Cas d'ús – L'obtenció de Notícies (I)

A aquest últim diagrama (*Figura 12*) d'aquesta secció dóna una visió de quines accions es poden portar a terme i de qui les realitza a cada moment. Aquest cas també necessita de dos actors, com al diagrama anterior, però aquí, les accions del sistema són les que donen sentit a la resta, ja que els descriptors i els mitjans de comunicació són necessaris per a que el cercador que es programi s'orienti per la xarxa i pugui recopilar de les pàgines web necessàries tot el contingut que es desitgi.

Fitxes dels casos d'ús associats al diagrama

Cas d'ús	Consulta el llistat de notícies
Versió	1
Descripció	Consultar el llistat de notícies.
Actors	Usuari.
Precondició	L'usuari està donat d'alta al sistema i té introduïts descriptors i associats mitjans de comunicació per poder iniciar els seguiments.
Flux Principal	1. Entrar a l'àrea del portal corresponent. 2. Visualitzar el llistat de notícies. a. Pot filtrar segons el mitjà de comunicació. b. Pot filtrar segons el descriptor que desitgi.
Postcondició	Llistat de notícies consultat.

Taula 11: Fitxa – Consulta el llistat de notícies

Aquesta fitxa reflecteix un dels procediments més habituals que realitzarà l'usuari al sistema. Per a que es pugui portar a terme correctament, és necessari que prèviament hagi introduït al sistema els paràmetres de cerca corresponents. És a dir, cal que hagi donat d'alta algun descriptor i que hagi escollit a quin o quins mitjans de comunicació vol realitzar el seguiment. D'aquesta manera, l'usuari pot fer una selecció de les notícies que desitja llistar i fer una cerca de la notícia més acurada.

Cas d'ús	Consulta d'una notícia
Versió	1
Descripció	Consultar el detall d'una notícia.
Actors	Usuari.
Precondició	L'usuari té alguna notícia associada al sistema.
Flux Principal	1. Entrar a l'àrea del portal corresponent. 2. Visualitzar el llistat de notícies . 3. Escollir la notícia que es vol veure el detall.
Postcondició	Notícia consultada.

Taula 12: Fitxa – Consulta d'una notícia

Aquesta fitxa pot donar-se a continuació de l'anterior, és a dir, l'usuari veu el llistat de notícies i selecciona la que vol visualitzar el detall. D'aquesta manera accedeix a la informació concreta que es guarda a la Base de Dades sobre la notícia en particular, dades com el titular, la entradeta i el text; i més particularment característiques com a quin mitjà pertany, la data de la captura i l'enllaç d'on s'ha trobat la notícia. També podrà visualitzar quin o quins han estat els descriptors pels quals ha estat seleccionada la notícia per emmagatzemar al sistema i d'aquesta manera tenir una visió clara de perquè i d'on s'ha recopilat la notícia.

Cas d'ús	Recaptació de notícies
Versió	1
Descripció	Recol·lectar les notícies de les webs dels mitjans de comunicació.
Actors	Sistema.
Precondició	S'han trobat els patrons per la cerca i el sistema està connectat a la xarxa.
Flux Principal	1. Descarregar el codi font de la pàgina principal del mitjà de comunicació 2. Extreure els enllaços corresponent a notícies. 3. Descarregar el codi d'una notícia del mitjà de comunicació. 4. Filtrar si s'ha d'emmagatzemar (segons apareixen els descriptors, o no) 5. Si apareixen els descriptors, extreure la informació necessària de la notícia. a. Extreure el titular b. Extreure l'entradeta c. Extreure el text 6. Emmagatzemar al sistema amb les relacions pertinents.
Postcondició	Descriptor esborrat del sistema

Taula 13: Fitxa – Recaptació de notícies

Per últim, es troba la fitxa de recaptació de notícies. L'actor d'aquesta fitxa és el Sistema i és l'encarregat d'analitzar les pàgines emmagatzemades a la Base de Dades i fer la recopilació dels enllaços cap a on es localitzen les notícies. En el moment de localitzar-les es passa el corresponent filtre per esbrinar si contenen la informació necessària i de ser així, analitza l'estructura per poder emmagatzemar la informació que es desitja.

7.3 Identificació de Subsistemes d'Anàlisi (v.1)

7.3.1 Identificació i Definició de Subsistemes (v.1)

A continuació es procedirà a identificar els subsistemes que formen part del projecte que s'està desenvolupant. Aquests subsistemes vindran donats pels requisits especificats en els punt anteriors, de tal manera que els solucionaran i quedaran repartits d'una manera semblant a la vista en el punt 7.1.2.

Així es poden trobar tres subsistemes diferenciats, on es toquen àrees diferents del projecte. El detall de cada es passa a comentar a continuació i quedarà reflectit gràficament a la *Figura 13* (a la pàgina següent) amb el Diagrama de Flux de Dades de Nivell 1.

Descriptors

A aquest subsistema es gestiona tot el relacionat amb els ítems o descriptors que l'usuari vol utilitzar per fer els seguiments de premsa. Així, les dades s'emmagatzemen a la Base de Dades fent de la gestió de l'àrea un recurs senzill i pràctic per l'ús de l'aplicació en qualsevol moment que l'usuari desitgi.

Mitjans de Comunicació

Tal com succeeix al subsistema de Descriptors, aquest subsistema també està encarat per a que s'utilitzi per orientar la cerca per la Xarxa. L'usuari administra els mitjans de comunicació, dels quals vol extreure les notícies que li interessen. D'aquesta manera, el sistema que prèviament a establert a la seva algorísmica els patrons necessaris per extreure aquesta informació buscarà per les pàgines d'aquests mitjans les notícies que coincideixin amb els descriptors que l'usuari a configurat prèviament.

Notícies

El subsistema de notícies és el més complex dels tres, ja que és l'encarregat de recopilar tota la informació de la Xarxa, analitzar-la i emmagatzemar-la (si cal) a la Base de Dades. D'aquesta manera, l'usuari pot trobar a l'àrea reservada de l'aplicació tot el recull de notícies que el sistema a recopilat per a ell, seguint els paràmetres que ell mateix a establert.

Cal tenir en compte que al subsistema de notícies hi ha dos àrees independents on una està dirigida pel sistema, essent l'encarregat de la cerca per la xarxa i exercint d'*spider* per les diferents pàgines dels mitjans de comunicació i l'altre àrea del subsistema vindrà regida per l'usuari, on podrà visualitzar de diverses maneres tota la informació que el sistema recopili per a ell.

Diagrama de Flux de Dades de nivell 1

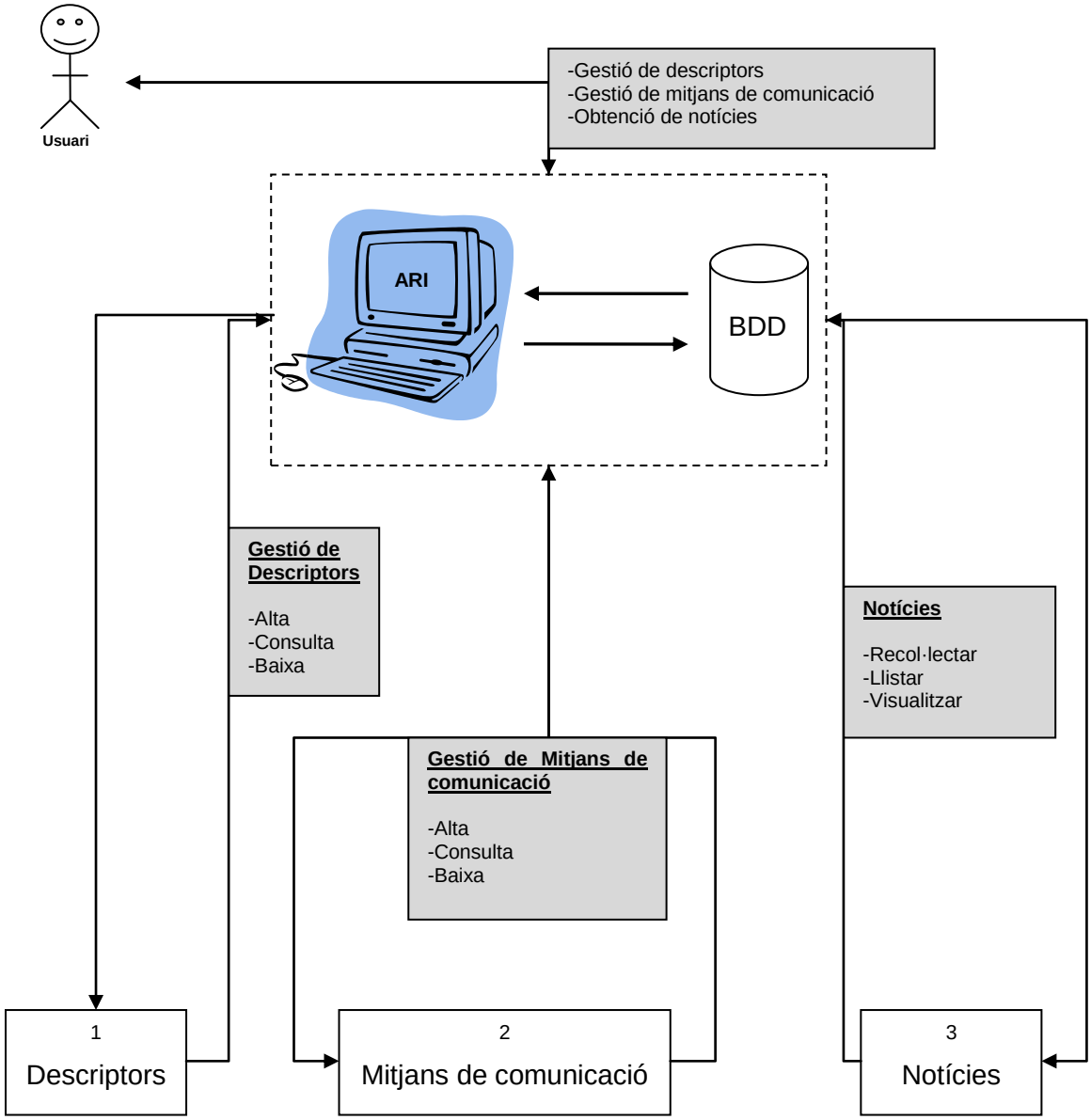

Figura 13: Diagrama de Flux de Dades de Nivell 1 (I)

7.4 Elaboració del Model de Dades (v.1)

7.4.1 Model Lògic de Dades Sol·licitat (v.1)

Una vegada estan localitzats els requeriments i les necessitats que ha de solucionar el projecte que s'està desenvolupant; i després de tenir clara la divisió de quins són els subsistemes adequats pel desenvolupament òptim del treball, s'ha de passar a identificar les entitats que es troben dins de l'àmbit del sistema d'informació, detallant els atributs de cada una d'elles i diferenciant clarament quins d'aquests són els més importants de cada entitat comentada.

Per a arribar a aquest propòsit, es passa a mostrar gràficament el diagrama de classes (*Figura 14*), per a continuar amb un disseny lògic on s'especifica més detalladament dels diferents taules i camps amb les quals està composta la Base de Dades de l'aplicació que s'està desenvolupant.

Diagrama de Classes

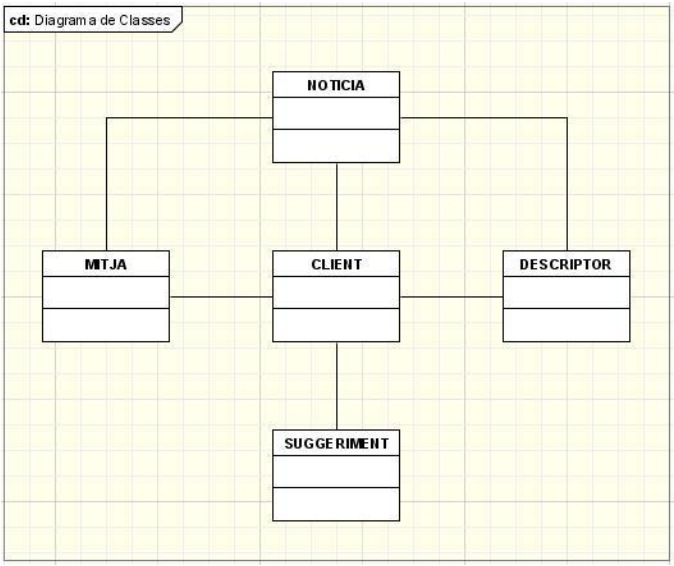

Figura 14: Diagrama de Classes (I)

Observant el diagrama de classes (*Figura 14*), que recolza el sistema, es pot deduir que la senzillesa és la característica principal. Tota l'aplicació gira al voltant de dues classes, CLIENT i NOTICIA. Aquests classes són les que porten tot el pes dels processos i les que determinen el que l'usuari acabarà observant al sistema quan hi accedeixi.

Tot seguit, respectant aquesta estructura de classes, es mostraran els diferents models de processos que existeixen al llarg de cada subsistema de l'aplicació.

7.5 Elaboració del Model de Processos (v.1)

Tal com s'ha pogut observar a l'apartat 7.3.1 (Identificació i definició dels Subsistemes), s'han identificat 3 subsistemes d'anàlisis. A continuació es farà la presentació dels corresponents models de processos que clarificaran tot el sistema i donaran una visió simple i clara de com ha de portar-se a terme l'execució de la programació per obtenir els objectius establerts.

Per arribar a tenir aquesta visió més precisa dels processos que es localitzen a cada subsistema es mostraran un seguit de gràfics amb els diagrames de flux de nivell 2 a on es presenten cada subsistema del punt 7.3.1.

7.5.1 Gestió dels Descriptors (v.1)

És el subsistema encarregat de gestionar els descriptors realitza les tasques de donar d'alta un ítem i facilitar a l'usuari la seva posterior consulta o eliminació. Els processos que es troben dintre d'aquest subsistema venen relacionats amb l'obtenció o inserció d'informació a la Base de Dades. De tal manera, cal que aquest processos siguin el més transparents possibles per a l'usuari i que pugui realitzar les accions d'aquesta àrea sense haver de saber res de com està muntat tot el sistema d'informació.

A continuació es detallarà cada procés:

Procés 1: Donar d'alta un Descriptor.

Per poder donar d'alta un descriptor al sistema, se li ha de facilitar a l'usuari un procés a on se li especifiqui quines són les dades necessàries perquè el cercador, posteriorment, pugui fer la recaptació per la xarxa. Així doncs, s'ha decidit mostrar a l'usuari a la seva àrea de client de l'aplicació un formulari d'entrada de dades. Aquestes seran les següents:

- *Nom del descriptor*: paraula o conjunt de paraules que voldrà trobar a les notícies que vol que el sistema recopili.
- *Descripció del descriptor*: petit text explicatiu que recolzarà el sentit de la paraula o paraules introduïdes al sistema de cerca.
- *Llistat de descriptors ja introduïts al sistema:* llistat estadístic amb el nombre de notícies associades a cada descriptor, que dóna la possibilitat a l'usuari d'associar-lo als seus seguiments.

Procés 2: donar de baixa un Descriptor

Aquest procés necessita que l'usuari especifiqui quin o quins descriptors vol anular del sistema de cerca. Per això, a l'àrea de gestió de descriptors del sistema s'oferirà un procés a l'usuari que el permeti seleccionar els ítems que desitgi per la seva posterior eliminació del sistema.

Aquest procés pot ser de dos tipus:

- *Eliminació parcial*: aquesta vindrà donada quan més d'un usuari tingui donat d'alta el mateix descriptor, que farà que el procés només elimini de la Base de Dades la relació entre el descriptor i l'usuari, reflectida a la taula CLIENT_DESCRIPTOR. I també en el cas de que hi hagin notícies al sistema que s'hagin recopilat gràcies a aquell i aparegui la seva relació a la taula DESCRIPTOR_NOTÍCIA.
- *Eliminació total*: L'eliminació total del descriptor es donarà quan a la Base de Dades no existeixi cap relació entre el descriptor i cap més client que l'usuari que està realitzant el procés de baixa i no existeixi cap notícia al sistema que s'hagi recaptat amb l'ajuda d'aquest descriptor. Només si es donen aquest dos casos, el descriptor serà eliminat de la taula DESCRIPTOR de la Base de Dades.

Procés 3: Consulta d'un Descriptor

Realitzar la consulta dels descriptors és un procés que es realitza automàticament el sistema quan accedeix a la seva àrea de gestió de descriptors de l'aplicació. De tal manera, el procés s'executa automàticament quan l'usuari accedeix a aquesta zona.

Així doncs, el procés de consulta d'un descriptor vindrà automatitzat pel sistema i oferirà a l'usuari un llistat dels descriptors que té associats i que li indicaran quines són les característiques per les quals el sistema li recopilarà notícies.

A continuació es visualitzarà el diagrama de flux de dades de nivell 2 (*Figura 15*), corresponent a aquest subsistema i els seus processos:

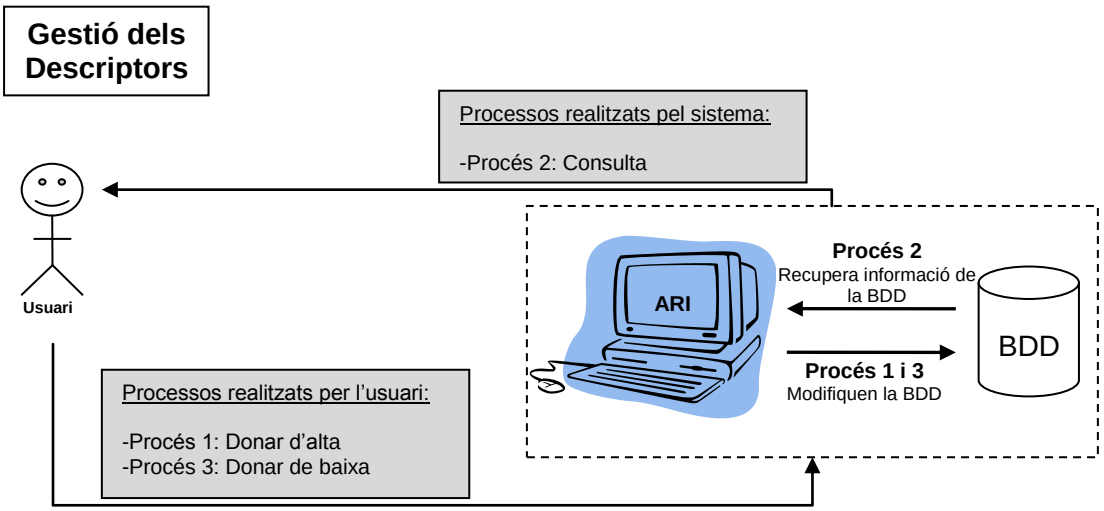

Figura 15: Diagrama de Flux de Dades de Nivell 2 – Subsistema de Gestió dels Descriptors

7.5.2 Gestió dels Mitjans de Comunicació(v.1)

Aquest subsistema, com ja s'ha pogut observar al punt d'identificació (el 7.1.3), és molt semblant a nivell de processos que el comentat prèviament, sempre i quan es visualitzi des del punt de vista de l'usuari. Així doncs, es troben els processos de donar d'alta, de baixa i consultar el llistat de mitjans de comunicació relacionats amb l'usuari. El punt de diferencia es troba a que l'usuari només pot donar-se d'alta en una sèrie acotada de mitjans de comunicació i que és el sistema el que determina aquest llistat a mida que s'incorporin a la Base de Dades.

A continuació es mostra el detall de cada procés que es troba dins d'aquest subsistema:

Procés 1: donar d'alta un Mitjà de Comunicació.

En aquesta versió es troba que el sistema ha de donar d'alta el mitjà de comunicació a la Base de Dades per a que l'aplicació pugui fer la cerca. Aquest procés el dur a terme l'administrador del sistema que inclourà a la Base de Dades la informació necessària del nou mitjà de comunicació que es pugui donar d'alta: nom, descripció, enllaç i si està actiu o no.

Procés 2: donar de baixa un Mitjà de Comunicació

Aquest procés també el realitza el sistema, l'administrador, que NO esborra de la Base de Dades la informació, sinó que modifica el camp MITJA_ACTIU i el desactiva per a que posteriors usuaris no puguin donar-se d'alta i els usuaris actuals no rebin més notícies d'aquest mitjà encara que estiguin a la seva llista.

A l'igual que succeeix amb la Gestió de Descriptors l'eliminació del sistema es pot donar de dos tipus:

- *Eliminació parcial*: aquesta vindrà donada quan una notícia o més estiguin relacionades amb el mitjà que es vol donar de baixa. Així doncs, si això succeeix, només es canviarà l'atribut de la taula MITJA per a que no es puguin donar d'alta més usuaris i el sistema no l'utilitzi en la seva cerca per la xarxa.
- *Eliminació total*: L'eliminació total del mitjà de comunicació es donarà quan a la Base de Dades no existeixi cap relació entre el mitjà i les notícies que hi ha al sistema. Quan aquestes condicions es donin, s'esborra de la Base de Dades la informació del mitjà de comunicació.

D'aquesta manera, el sistema pot assegurar-se la consistència de les dades i evitar que l'usuari a l'utilitzar el sistema es trobi amb incongruències.

Procés 3: donar-se d'alta a un Mitjà de Comunicació

L'usuari pot realitzar el procés de donar-se d'alta a un mitjà de comunicació un cop el sistema ha donat d'alta al menys a un mitjà. Així, anirà a l'àrea corresponent de l'aplicació i es trobarà amb un desplegable amb el llistat dels possibles mitjans que pot afegir per portar a terme els seguiments. El procés de donar-se d'alta finalitzarà quan l'usuari hagi escollit el mitjà i hagi realitzat l'acció de donar-se d'alta mitjançant el botó que facilita l'aplicació al sistema.

Procés 4: donar-se de baixa d'un Mitjà de Comunicació

El procés de donar-se de baixa d'un mitjà de comunicació ho realitza l'usuari quan no vol rebre més notícies d'un determinat mitjà. D'aquesta manera, anirà a la zona reservada per la gestió dels mitjans de comunicació i seleccionarà el medi corresponent que vol donar de baixa del seu sistema gestor de seguiments de premsa.

Procés 5: Consultar un Mitjà de Comunicació

Per últim procés d'aquest subsistema es troba la consulta dels mitjans de comunicació als quals està subscrit l'usuari. Aquest procés és molt similar al comentat prèviament a la Gestió de Descriptors. Aquest és un procés que fa el sistema automàticament quan l'usuari entra a la zona de gestió de mitjans de comunicació, facilitant a l'usuari l'accés al detall dels mitjans que fins al moment té contractat seguiments de premsa.

Per deixar-ho més clar, a continuació es visualitzarà el diagrama de flux de dades de nivell 2 (*Figura 16*), corresponent a aquest subsistema i els seus processos:

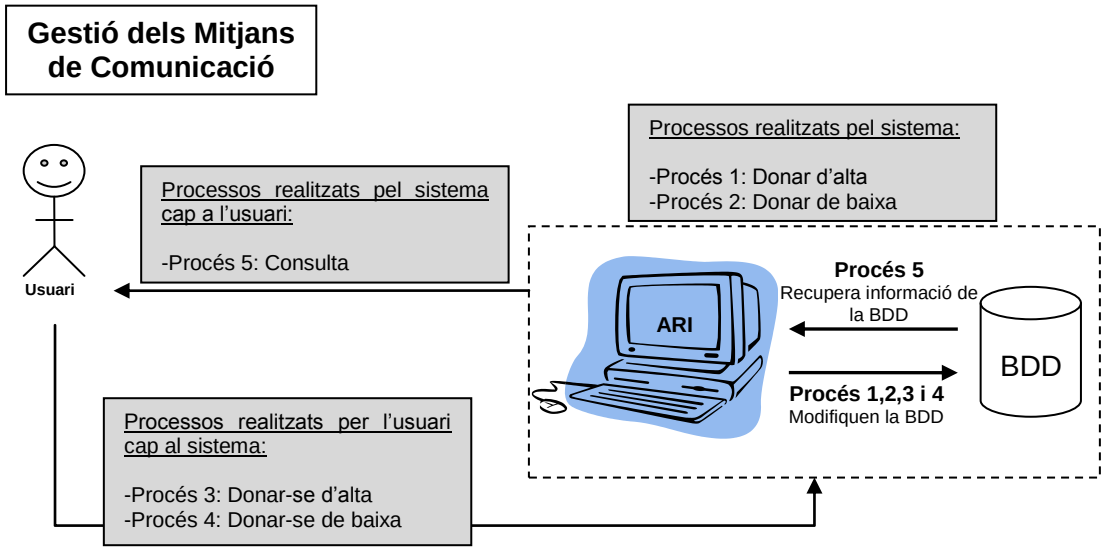

Figura 16: Diagrama de Flux de Dades de Nivell 2 – Gestió dels Mitjans de Comunicació

7.5.3 Gestió de Notícies (v.1)

Aquest subsistema es diferencia dels altres dos subsistemes comentats en que els seus processos són els encarregats d'obtenir la informació d'Internet. De la mateixa manera que passa amb el subsistema de Gestió dels Mitjans de Comunicació, hi ha dos tipus de processos diferenciats, els que són executats pel sistema i els que realitza l'usuari. La principal diferència entre aquest subsistema i els dos presentats anteriorment, és que els processos més importants són els que realitza el sistema.

A continuació es mostrarà el detall dels tres processos que formen aquest subsistema:

Procés 1: recopilar Notícies d'Internet

Aquest procés és el més important, no només d'aquest subsistema sinó, de tot el projecte. Aquest procés està comandat pel sistema i és l'encarregat d'anar per les pàgines web dels mitjans de comunicació escollits i buscar les notícies que corresponguin als descriptors que es desitja trobar a cada moment.

Per poder fer això, el sistema ha de portar a terme un seguit de passos que es comenten:

1. *Recuperar enllaços dels Mitjans de Comunicació*: cal qui el sistema vagi a la Base de Dades i recuperi la direcció web, l'enllaç, a on es pot trobar el contingut de cada mitjà.
2. *Descarregar el contingut de l'enllaç*: cal que el sistema descarregui la informació de la pàgina web i trobi tots els enllaços. Cal que faci una selecció i es quedi amb només els enllaços que corresponen a notícies, eliminant tots els que facin referència a publicitat, altres seccions, contactes, ...
3. *Recuperar els descriptors que s'han d'utilitzar per la selecció*: el sistema ha d'anar a la Base de Dades i recuperar els descriptors que s'han d'utilitzar a la cerca del mitjà de comunicació que s'està extraient els enllaços de notícies.
4. *Descarregar el contingut dels enllaços trobats*: el sistema ha d'anar enllaç per enllaç obtingut al pas anterior i filtrar pels descriptors. Un cop es filtra poden donar-se dos vies. Aquest procés es detallarà en profundament en punts posteriors de la documentació (al capítol 7.8)
5. *Verificació de coincidències*
    a. *No hi ha coincidència entre els descriptors i el contingut de la notícia*: es passa a l'enllaç següent i el sistema guarda l'enllaç per no consultar-ho en posteriors cerques.
    b. *Hi ha coincidència entre els descriptors i el contingut de la notícia*: ha d'analitzar el contingut de la notícia i localitzar les parts principals de les quals es composa.
6. *Emmagatzemar a la Base de Dades el resultat*: si el sistema ve del punt 5.a. llavors només emmagatzemarà a la taula DESCARTS l'enllaç corresponent. En canvi, si la línia del procés ve donada pel 5.b. llavors el sistema cal que emmagatzemi les parts de la notícia a la taula NOTICIA i les relacions pertinents a les taules MITJA_NOTICIA i DESCRIPTRO_NOTICIA per tenir enllaçada tota la informació per la seva posterior consulta. També, de la mateixa manera que succeeix amb els enllaços que no contenen la informació desitjada, un cop s'emmagatzema la informació l'enllaç de la notícia recaptada també s'afegirà a la taula DESCARTS per a que no es consulti en posteriors escombrats de la xarxa.

Aquest procés és un procés iteratiu que l'administrador del sistema ha de decidir quan es llança, si ho fa manualment o automàticament, cada quan es vol fer escombrats dels mitjans de comunicació (cada hora, diaris, setmanals...). Aquestes decisions es prenen a mida que es va analitzant els mitjans de comunicació i es va veient la seva evolució i actualització de la informació que aporten, cada quan publiquen novetats o cada quan fan neteja de les notícies velles.

Procés 2: llistar Notícies

El procés de llistar notícies és un procés molt semblant als comentats a la Gestió de Descriptors i a la Gestió de Mitjans de Comunicació. Així doncs, és gairebé igual al procés de consulta d'un descriptor o d'un mitjà de comunicació. És un procés automàtic que facilita el sistema a l'usuari quan aquest accedeix a l'àrea corresponent del projecte, però, com a diferència amb els dos subsistemes anteriors, cal dir que l'usuari té part d'interacció amb el sistema ja que se li dona la possibilitat de modificar el llistat que veu.

L'usuari, quan accedeix al sistema i va a l'àrea de gestió de notícies hi troba un llistat de totes les notícies que s'ha recopilat per ell, ordenades per ordre de l'última recopilada a la primera.

D'aquesta manera sempre veu les últimes novetats trobades pel sistema. També, a aquesta àrea se li dóna la possibilitat de filtrar les notícies de la llista segons el mitjà de comunicació o el descriptor que ell desitgi, donant com a resultat una versió acotada del llistat principal.

Procés 3: consultar el detall d'una Noticia

Aquest procés el realitza l'usuari. Un cop a visitat el llistat, el sistema li dóna la possibilitat de veure el detall de la notícia. Porta a terme aquesta opció podrà veure el que el sistema té emmagatzemat sobre la notícia escollida, totes les relacions amb els descriptors que han fet que es descarregui de la xarxa i l'origen de la mateixa.

Per últim, per deixar-ho mes clar, a continuació es visualitzarà el diagrama de flux de dades de nivell 2, *Figura 17*, corresponent a aquest subsistema i els seus processos:

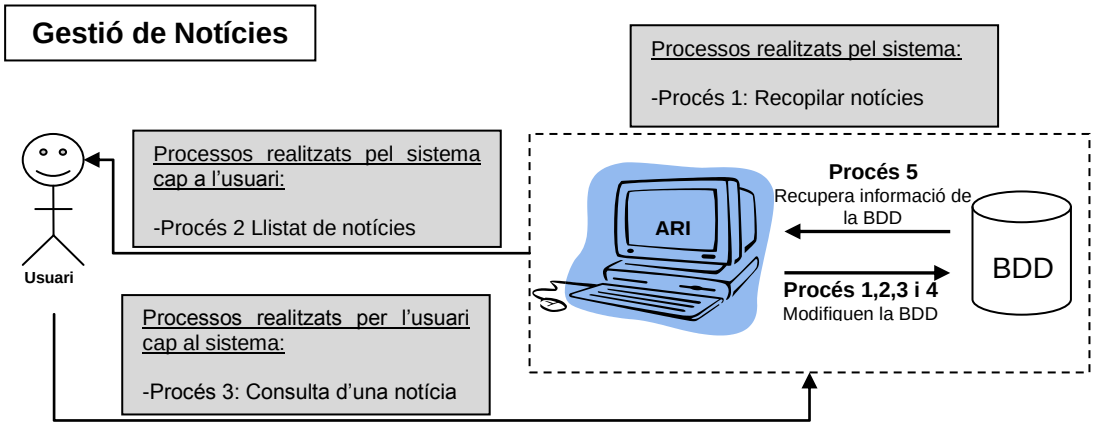

Figura 17: Diagrama de Flux de Dades de Nivell 2 – Gestió de Notícies

7.6 Definició d'Interfícies d'Usuari (v.1)

7.6.1 Especificació de Principis Generals de la Interfície (v.1)

Per a poder fer una bona especificació dels principis generals de la interfície es passarà a mostrar els dissenys dels prototips de les pàgines que utilitzarà l'usuari mentre estigui dintre del sistema. Es podrà observar com les línies de l'aplicació són senzilles per fer la seva utilització més amena i atractiva per l'usuari.

D'aquesta manera s'estableix que totes les pantalles que visualitzi l'usuari hauran de seguir unes característiques generals i unes línies bàsiques de disseny. A continuació es mostrarà gràficament uns esquemes on es podran veure reflectits els diferents tipus de pantalles que es trobaran al llarg del sistema que es vol implementar.

Per implementar aquest disseny s'han utilitzat fulles d'estil [11] per determinar l'homogeneïtat entre les diferents pàgines de l'aplicació.

Pàgina Principal

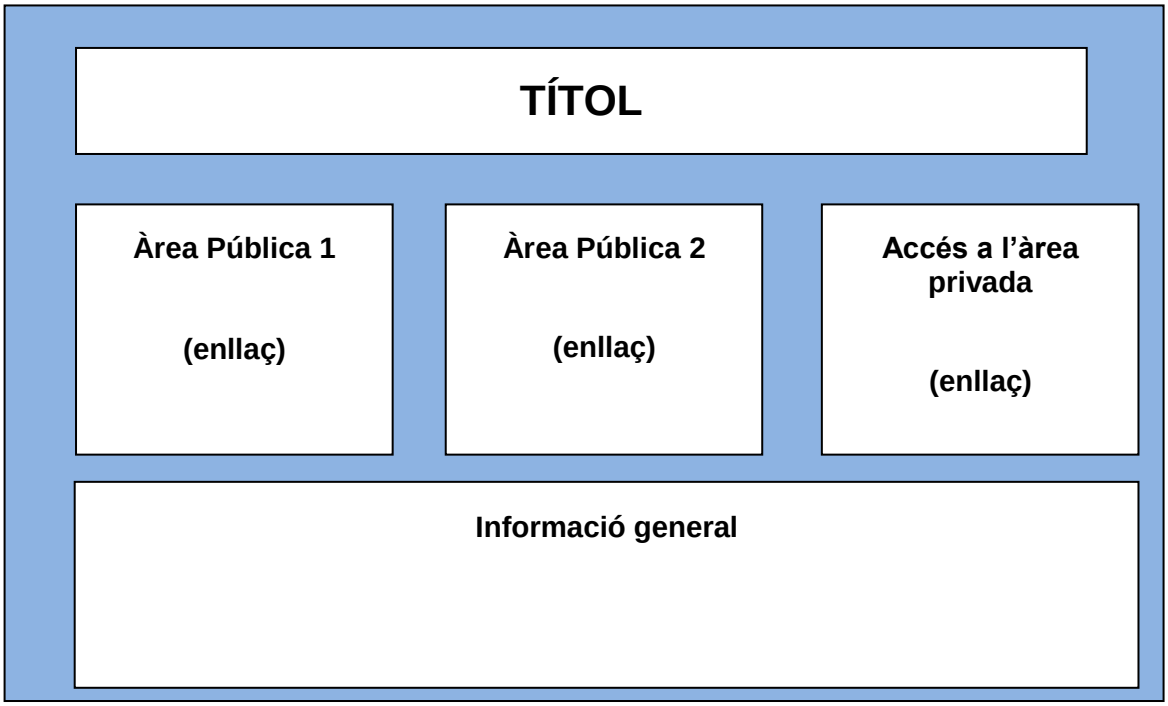

Figura 18: Esquema de la interfície – Pàgina Principal

Tal com es pot observar a la *Figura 18*, a la pàgina principal es podrà localitzar els accessos a les diferents àrees de l'aplicació, tant a l'àrea pública com a l'àrea privada.

A la zona d'informació general es localitzaran les dades sobre el que l'espera a l'usuari al llarg de la seva investigació per l'aplicació. Una introducció al món del *press-clipping* i un històric de dates a on es reflecteixen els diferents esdeveniments que puguin anar succeint al llarg del desenvolupament del projecte, ja que com es pot entreveure amb la metodologia de treball escollida, l'aplicació anirà evolucionant a mida que es vagi investigant més sobre els temes que el projecte abasta.

La mobilitat vindrà donada pels enllaços que seran les zones delimitades com: àrea pública 1, àrea pública 2 i l'accés a l'àrea privada.

Pàgines Públiques

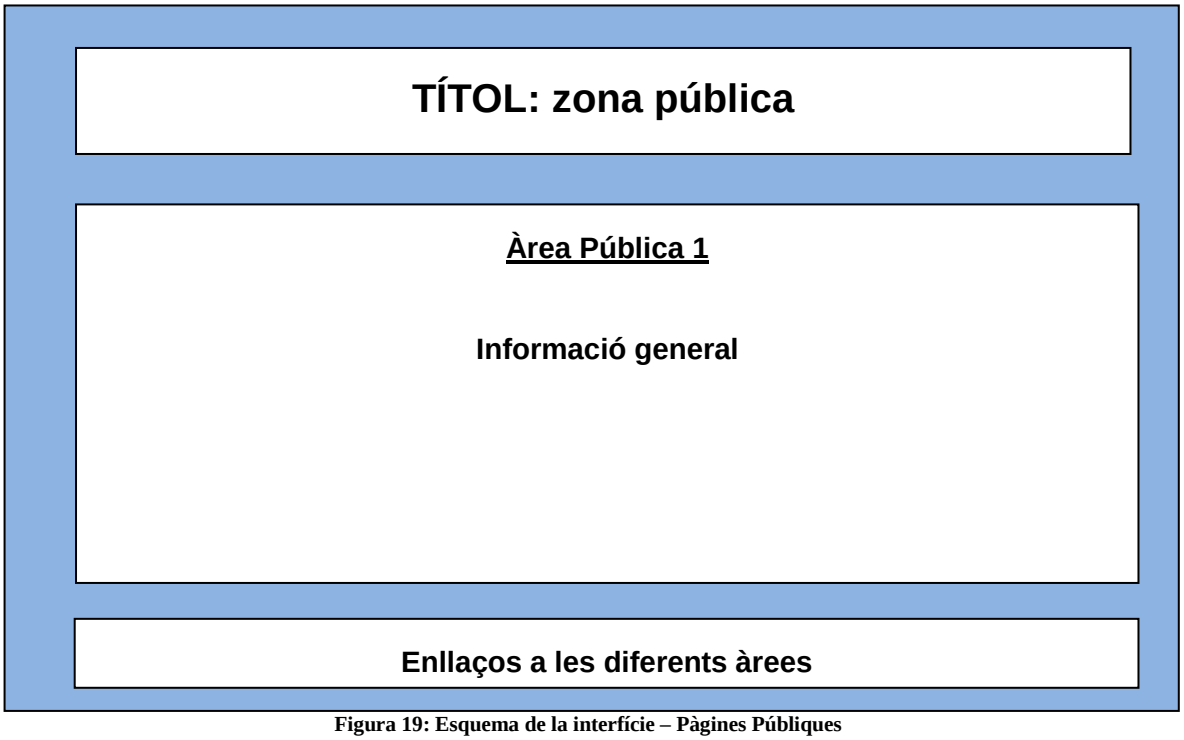

Figura 19: Esquema de la interfície – Pàgines Públiques

A les pàgines públiques, *Figura 19*, l'esquema que es visualitza vindrà donat per un espai ampli d'informació que ocuparà la part central de la pantalla. Així l'usuari es podrà centrar en la lectura sense obstacles.

L'estructura està dividida de la següent manera:

- A la part superior de la pantalla s'observarà el títol de la aplicació amb el detall de la secció pública a la que es troba l'usuari, això farà que mai perdi la consciència de què està veient. També, per fer molt més senzilla la utilització de l'eina.
- A la part de sota de la pantalla es localitzaran els enllaços que fan que l'usuari pugui navegar per tota l'aplicació.

Pàgines Privades

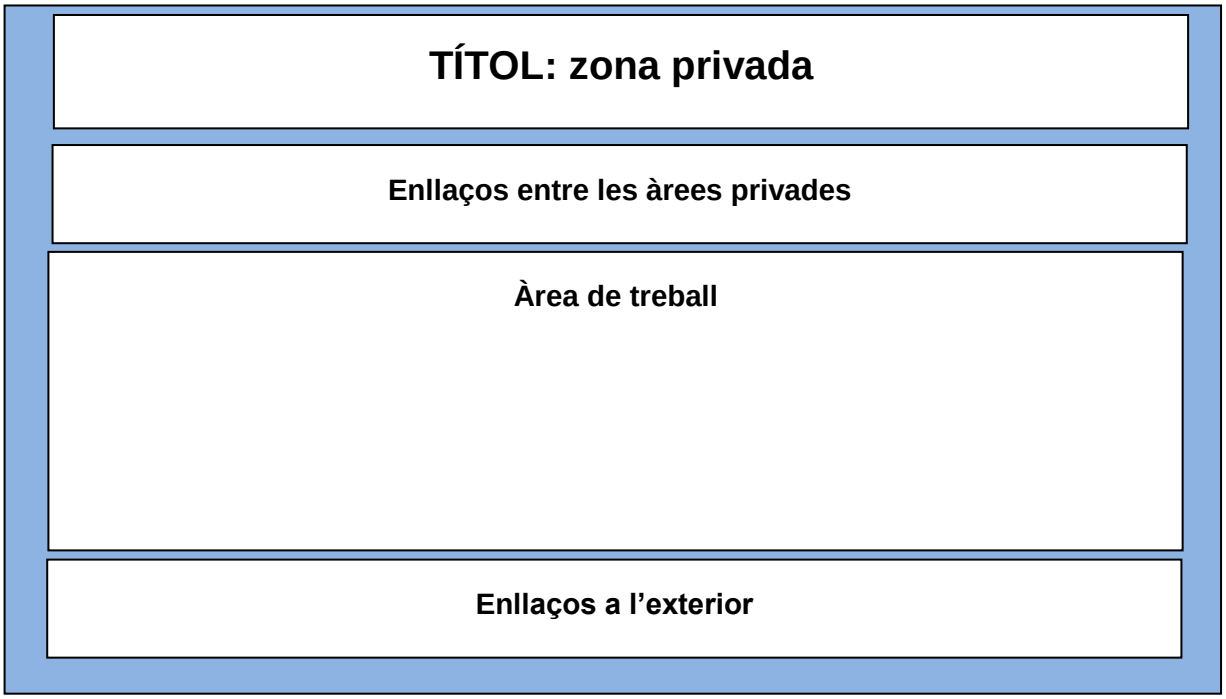

Figura 20: Esquema de la interfície – Pàgines privades

Aquest esquema reflecteix l'estructura bàsica de com estaran formades les pàgines de l'àrea privada (*Figura 20*). Hi haurà una zona central de treball delimitada superior i inferiorment per zones d'enllaços, la superior hi seran els enllaços per moure's per dins de la zona privada i la inferior serà per moure's cap a fora.

Al igual que succeeix als dos esquemes superiors, la part superior de la pantalla està reservada pel títol amb l'especificació de la conseqüent zona que es troba en cada moment l'usuari i quin treball pot realitzar en cada moment.

7.6.2 Identificació de Perfils i Diàlegs (v.1)

Després de l'explicació dels esquemes de les interfícies que es podran trobar al sistema, es passa a identificar les diferents finestres de l'aplicació i la seva jerarquia (*Figura 21*). Tal com es pot intuir, la jerarquia vindrà donada per una pàgina principal de la qual penjaran les ramificacions que podran pertànyer a pàgines públiques o pàgines privades.

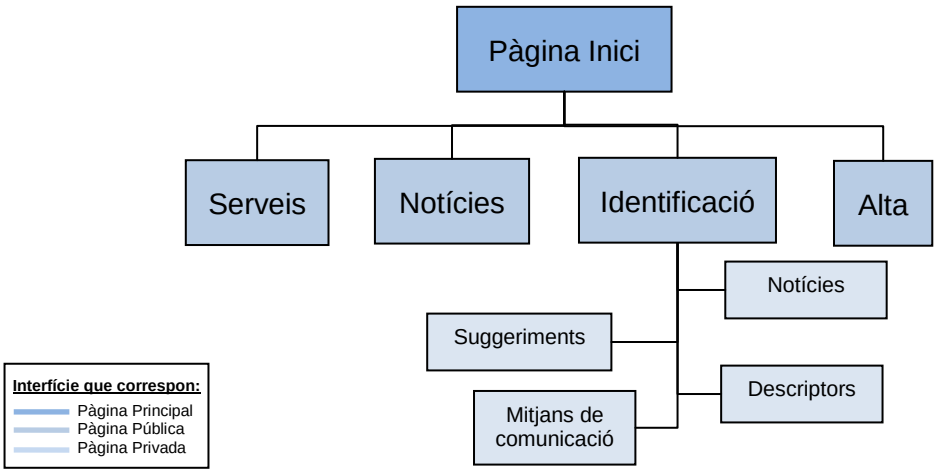

Figura 21: Jerarquia de pàgines de l'aplicació

7.6.3 Especificació de Formats Individuals de la Interfície de Pantalla (v.1)

Un cop es tenen en compte tots els requisits que s'han especificat pels usuaris participants i finals del sistema que aquí es planteja, s'ha passat al disseny de les interfícies gràfiques. De tal manera, s'han respectat les característiques i les necessitats establertes pels usuaris comentats al punt 6.2.2 d'aquesta memòria (Identificació dels Usuaris participants en l'Estudi de la Situació Actual).

A continuació s'ha optat per detallar el contingut de cada pantalla que pot trobar l'usuari, especificant totes les accions i dades que es necessiten per utilitzar cada àrea de l'aplicació i obtenir els resultats desitjats.

Pantalla d'inici

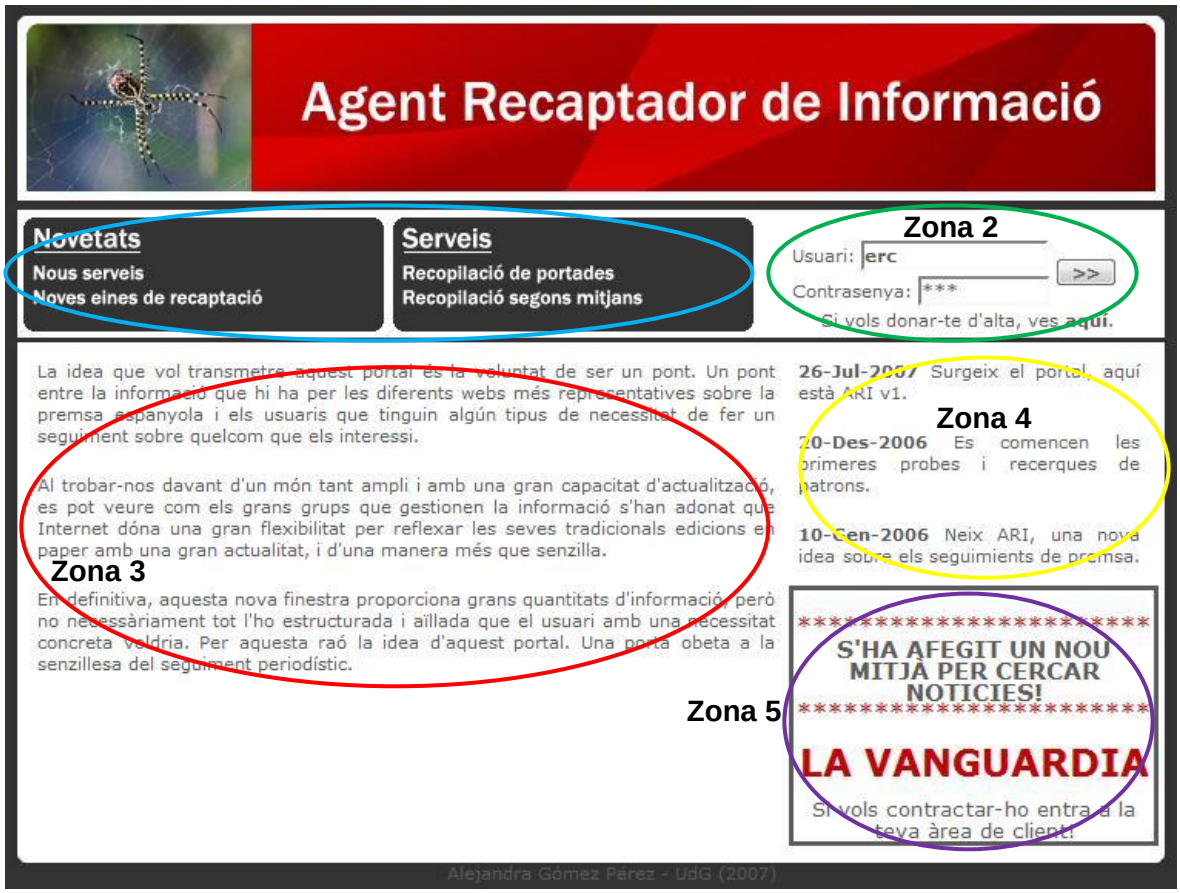

Figura 22: Pantalla d'inici de l'aplicació

A aquesta primera pantalla (*Figura 22*), es pot observar com serà la pàgina principal que trobarà l'usuari al iniciar-se al sistema.

Com a navegabilitat cal comentar que els blocs que es troben encerclats a la *Figura 22* són els enllaços que permetran a l'usuari moure's per les diferents àrees, tan públiques com privades. El cercle blau (zona 1) engloba el que serien els dos enllaços cap a pàgines públiques de l'aplicació i el cercle verd (zona 2) emmarca l'enllaç per donar-se d'alta com usuari del sistema i el formulari d'accés a l'àrea privada a on trobarà les diferents parts de gestió que es poden veure que pengen del mòdul d'identificació de la *Figura 21* : Descriptors, Mitjans de Comunicació, Notícies i Suggeriments.

A mode informatiu hi ha tres diferenciades:

- *Cercle vermell*: (zona 3) Informació general sobre la eina.
- *Cercle groc*: (zona 4) Històric de notícies relacionades amb el desenvolupament de l'aplicació.
- *Cercle lila*: (zona 5) Àrea reservada a un esdeveniment rellevant sobre l'aplicació.

Pantalla de Serveis

Figura 23: Pantalla de Serveis

La pantalla que a la *Figura 23* s'observa el detall dels serveis que ofereix l'aplicació. En un principi, l'oferta està reduïda a un sol servei, la recaptació de notícies per l'usuari seguint els paràmetres que hagi configurat ell mateix prèviament. Però tal com s'informa després, aquesta oferta es veurà incrementada a mida que els usuaris del sistema vagin aportant suggeriments i desitjos de serveis.

També es pot observar a la part inferior de la pàgina com es localitzen els enllaços per donar a l'usuari la navegabilitat necessària per l'aplicació.

Pantalla de Novetats

Figura 24: Pantalla de Novetats

Aquesta pantalla (*Figura 24*) està reservada per a que el sistema informi a l'usuari de les novetats que vagin sorgint al voltant de l'aplicació.

Aquestes novetats tan poden ser del funcionament intern de l'aplicació, com l'aparició d'un nou mitjà de comunicació per a que l'usuari pugui fer nous seguiments, com a nivell extern, respecte a la societat actual i el que influeix en l'aplicació per millorar els seus serveis.

Per últim, al igual que a la pàgina anterior, a la part inferior de la pàgina es troben situats els enllaços per la navegació per les diferents zones públiques de l'aplicació.

Pantalla D'Alta

Figura 25: Pantalla per donar-se d'alta un usuari

A continuació es passa a comentar el formulari d'alta que es pot observar a la *Figura 25* i que mostra totes les dades que necessita el sistema per poder enregistrar correctament una nova persona. D'aquesta manera, algú que vulgui passar a formar part de la comunitat d'usuari d'aquesta aplicació haurà d'aportar el següent:

- Nom i Cognoms
- Direcció
- Telèfon i Fax
- Correu electrònic
- Nom d'usuari i contrasenya

També haurà de decidir si vol rebre informació, via correu electrònic, de les novetats que vagin apareixent a mida que va evolucionant el sistema de captació de l'aplicació.

Tal com ve succeint a les dues pàgines anteriors, a la part inferior de la pàgina es localitzen els enllaços que permeten a l'usuari moure's per l'aplicació.

Pantalla d'Identificació

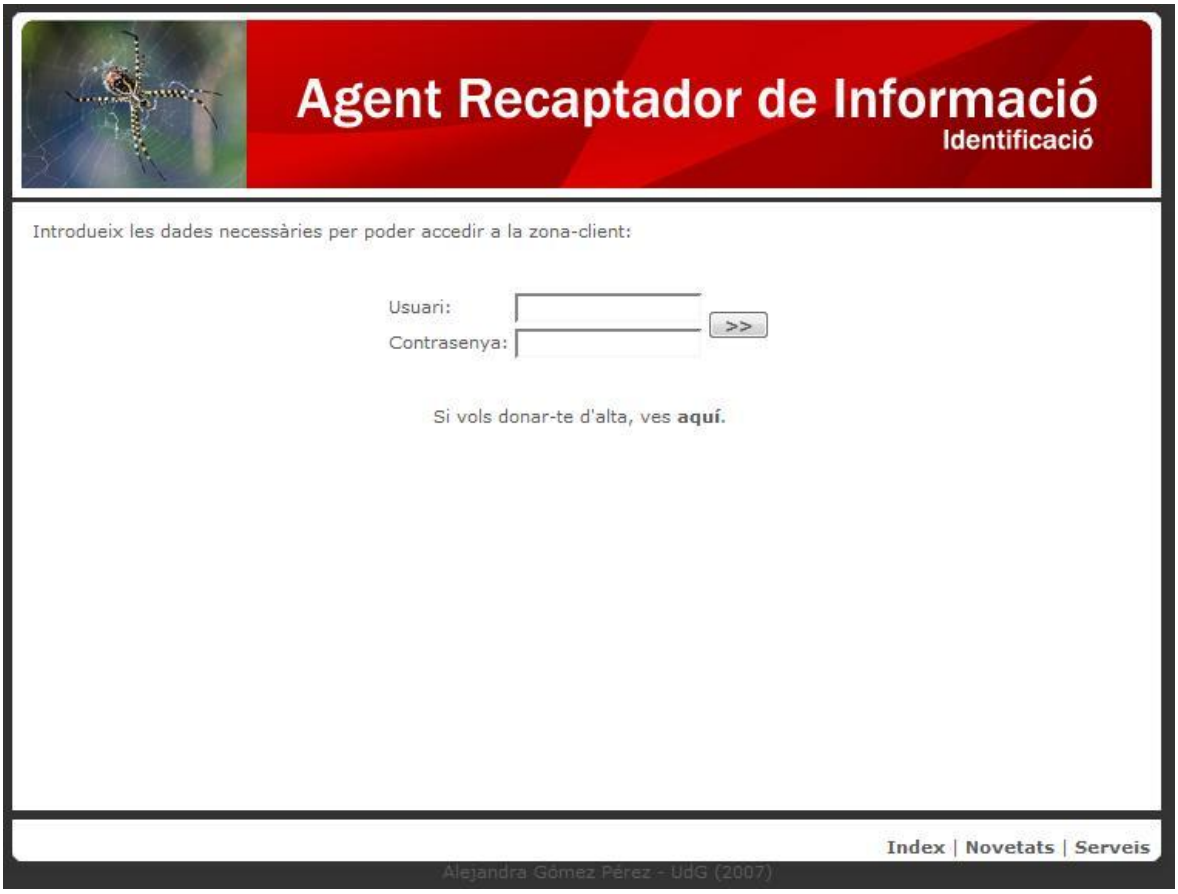

Figura 26: Pantalla d'Identificació d'accés a la zona privada

Com alternativa a la via d'entrada a l'àrea privada de la pàgina principal, es troba aquesta pantalla (*Figura 26*) que es pot arribar des de les pàgines públiques com la de serveis o novetats. Aquesta interfície ofereix a l'usuari un formulari d'entrada a l'àrea privada, alhora que facilita als usuaris no registrats al sistema un enllaç d'accés al formulari d'alta que s'ha explicat a la pàgina anterior.

Per últim, al igual que a totes les pàgines anteriors de la zona pública, a la part inferior de la pàgina es localitzen els enllaços necessaris per facilitar la navegació de l'usuari per la zona pública de l'aplicació.

Pantalla Principal de l'Àrea Privada

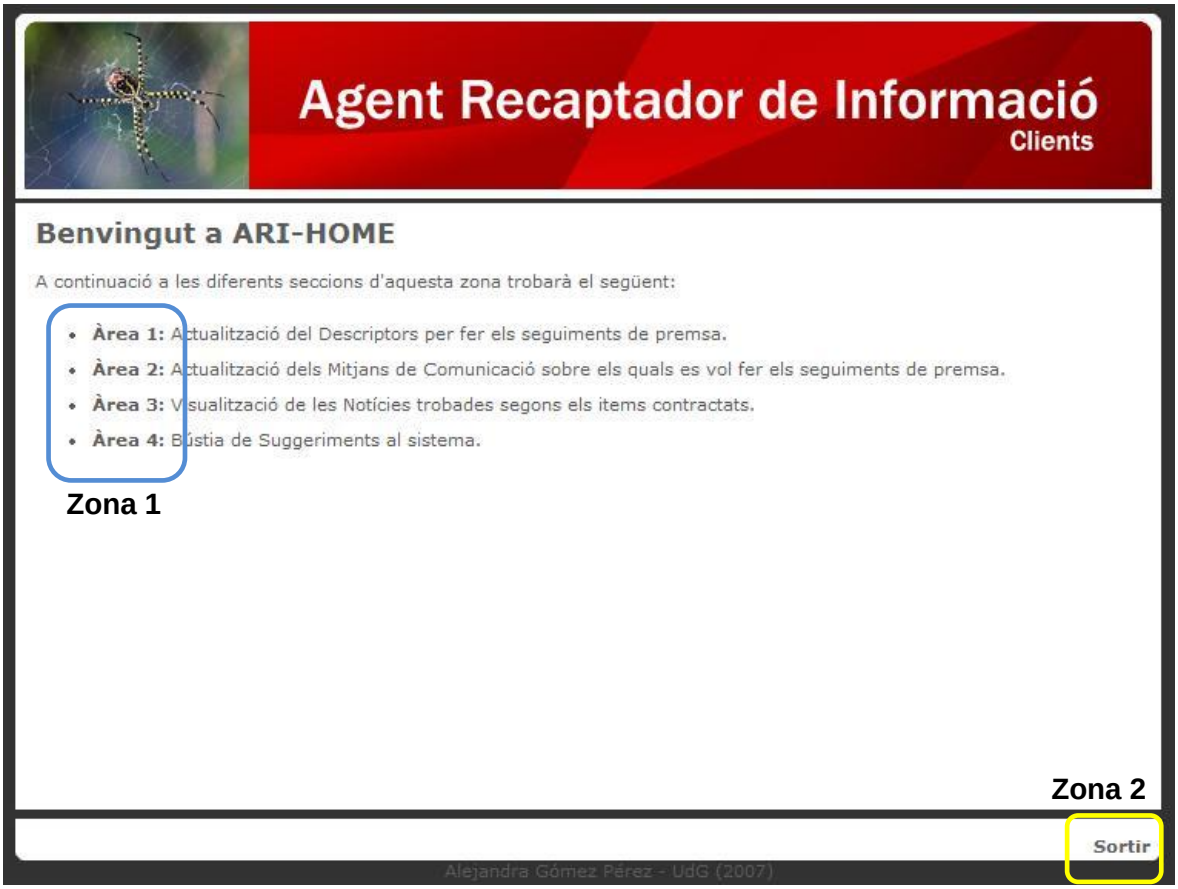

Figura 27: Pantalla Principal de la zona privada

Aquesta *Figura 27* mostra la pàgina d'entrada a l'àrea privada de l'usuari de l'aplicació. Aquí el que trobarà serà un resum de les diferents àrees que té el sistema i que ofereix diferents serveis a l'usuari. Així doncs, a la zona encerclada en blau (zona 1), el que correspondria a les lletres ressaltades amb negreta, són els diferents enllaços cap a les diverses àrees del sistema. Igualment, la zona ressaltada en groc (zona 2) delimita l'enllaç de sortida de la zona privada del sistema i que es repetirà al llarg de totes les pantalles que pertanyin a aquesta zona.

Pantalla de Gestió de Descriptors

Figura 28: Pantalla de Gestió de Descriptors

A la *Figura 28* que aquí s'observa es pot visualitzar la primera àrea de gestió del sistema. Aquesta, dona la possibilitat a l'usuari de gestionar els seus descriptors.

A la mateixa zona de treball hi trobarà dos formularis diferents:

- Primer: un llistat amb tots els descriptors que té contractats amb l'opció de seleccionar algun si el vol donar de baixa dels seus seguiments. En cas d'esborrar el sistema informarà de si s'ha realitzat bé l'acció i li facilitarà un enllaç per tornar a la pàgina de gestió.
- Segon: dos formularis per afegir nous descriptors al sistema.
    - El primer és un llistat amb els descriptors que ja estan incorporats al sistema i que l'usuari no té, amb el nombre de notícies que s'han recopilat per a ell.
    - El segon formulari amb la possibilitat d'incorporar nous descriptors.

    En tots dos casos, el sitema informarà a l'usuari del resultat de la seva acció i li facilitarà la tornada a la gestió principal. En cas d'arribar al límit de descriptors que el sistema deixa tenir a cada usuari, s'informarà tal com es pot observar a la *Figura 29*.

Figura 29: Informació de màxim número de Descriptors adquirit

Pantalla de Gestió de Mitjans de Comunicació

Figura 30: Pantalla de Gestió de Mitjans de Comunicació

Aquesta pantalla porta a terme la gestió dels mitjans de comunicació als que un usuari del sistema pugui estar subscrit. Així, a la *Figura 30* es pot observar com la pantalla està dividida també en dues zones, la superior es troba un llistat amb els diferents mitjans de comunicació dels quals l'usuari rep les notícies i que pot donar-se de baixa en qualsevol moment seleccionant-ho i prenent el botó d'esborrar. També, troba un llistat (a la imatge desplegat) a on es troben els mitjans de comunicació que el sistema analitzats i preparats per captar-hi notícies i que l'usuari encara no hi està donat d'alta. Per incorporar-ho als seus seguiments, l'usuari només ha de seleccionar el que vulgui i prémer el botó d'afegir.

Pantalla de Gestió de Notícies

Figura 31: Pantalla de Llistat de les Notícies

Aquesta àrea del sistema aporta a l'usuari la informació que vol obtenir al donar-se d'alta en un sistema com el que aquí es presenta. Aquesta àrea mostra un llistat de totes les notícies que el motor de cerca ha recopilat per a ell. Per facilitar la seva utilització s'han posar un parell de filtres, segons els dos paràmetres bàsics amb els que treballa l'usuari, sobre mitjans de comunicació i sobre descriptors. Així, l'usuari podrà obtenir diferents versions del llistat principal configurant aquests dos filtres.

Des d'aquesta pantalla es dona la possibilitat a l'usuari de visualitzar el detall de la notícia. Això ho podrà fer prenent al símbol + que hi ha al requadre de cada notícia del llistat. A continuació es veurà com es mostrarà el detall de cada notícia en particular.

Pantalla de Visualització d'una Notícia

Figura 32: Pantalla de Visualització d'una Notícia

A la *Figura 32* es pot veure el detall de la notícia que l'usuari hagi seleccionat a la pantalla anterior.

Aquí es pot visualitzar les diferents parts de la notícia:

- Titular
- Entradeta
- Cos

A sota de les dades pròpies de la notícia es troben els següents detalls:

- Mitjà de Comunicació d'on s'ha extret la notícia
- Data en que s'ha fet la captura
- Enllaç cap a la pàgina original on s'ha trobat

I per últim, al final de tot es troba un llistat dels descriptors que es localitzen a la notícia.

Pantalla de Suggeriments

Figura 33: Pantalla de Suggeriments

Aquesta última pantalla (*Figura 33*) és un pont entre l'usuari i l'administrador del sistema. Aquesta dóna la possibilitat a l'usuari d'enviar cap al sistema les seves peticions, recomanacions, observacions, ...

Així doncs, l'usuari a l'entrar a aquesta àrea trobarà una àmplia zona d'escriptura per poder esplaiar-se i finalment enviar-ho al sistema prenent el botó que es troba a sota de la zona.

7.7 Definició de l'Arquitectura del Sistema (v.1)

7.7.1 Definició de Nivells d'Arquitectura (v.1)

L'objectiu al que es vol arribar amb el desenvolupament d'aquest projecte és canviar l'actual sistema que té l'usuari per obtenir la recaptació de notícies que desitja. Per aconseguir aquest propòsit, aquest projecte aporta una arquitectura recolzada en la tecnologia web per a oferir a l'usuari un sistema que li aporti un seguiment de notícies senzill i eficaç per les seves necessitats.

Mitjançant el següent esquema (*Figura 34*) es visualitza com s'estructuren els diferents nivells.

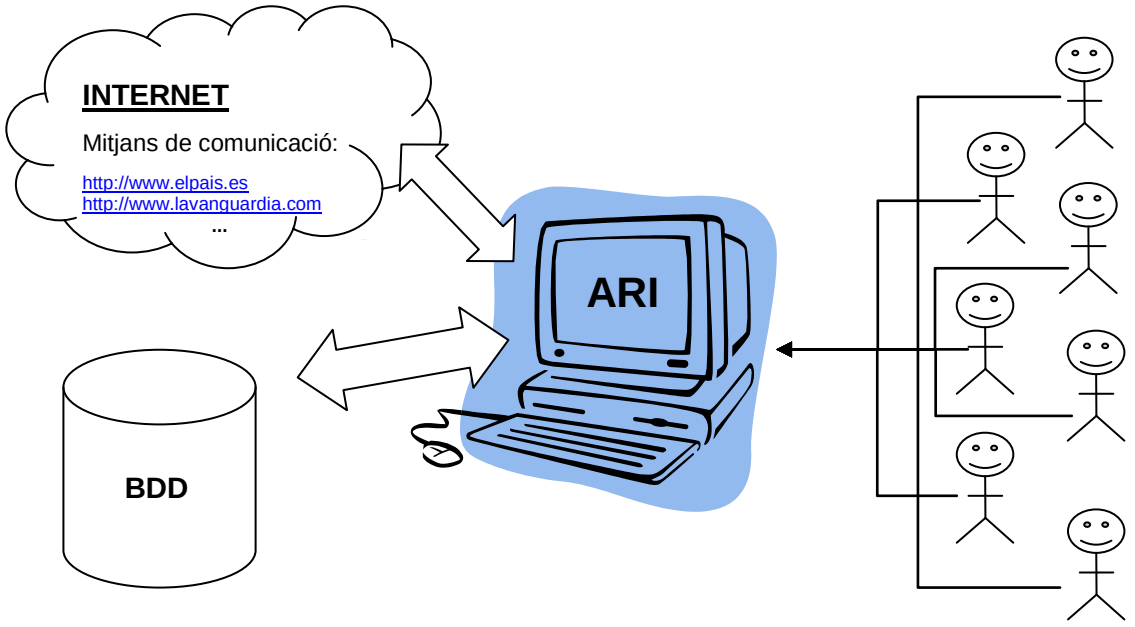

Figura 34: Esquema arquitectònic del sistema

Com es pot observar a la *Figura 34* els nivells en que està dividida l'arquitectura es basen sobretot en la separació entre el món d'Internet i el servidor del projecte a on es troba la Base de Dades i el sistema. Així, els usuaris connecten amb el sistema i aquest amb la xarxa a on anirà a buscar la informació, tal com l'hagi configurada l'usuari.

7.7.2 Especificació de l'Entorn Tecnològic (v.1)

L'aplicació que s'ha desenvolupat està allotjada a un entorn web. Així, els requisits tecnològics des del punt de vista del sistema administrador i els requisits tecnològics des del punt de vista de l'usuari seran els que es comenten a continuació:

- L'administrador del sistema que munta el projecte necessita un ordinador o servidor amb connexió a Internet, amb un sistema de gestió de Base de Dades relacional (*MySQL*) i un servidor HTTP per poder gestionar (*Apache*)
- L'usuari només necessitarà un ordinador amb connexió a Internet.

7.8 Disseny de l'Arquitectura del Sistema (v.1)

7.8.1 Disseny de Mòduls del Sistema (v.1)

A aquesta altura del desenvolupament del projecte es passa a dividir els subsistemes, declarats als punts anteriors, en mòduls que especifiquin l'estructura de l'aplicació. Per assolir aquest objectiu es mostrarà gràficament mitjançant diagrames d'estructura i mitjançant pseudocodi per especificar la lògica interna.

Després de l'anàlisi dels possibles mòduls en els que es podrien dividir els diferents subsistemes en els que s'ha anat dividint el projecte, s'ha arribat a la conclusió que, per la similitud en el funcionament dels processos que es porten a terme, sobretot en els subsistemes de Gestió de Descriptors i Gestió de Mitjans de Comunicació, s'especificarà a continuació aquest mòduls d'una forma genèrica i més profundament els processos referents al subsistema d'Obtenció de Notícies.

Mòduls del Sistema des del punt de vista de l'usuari

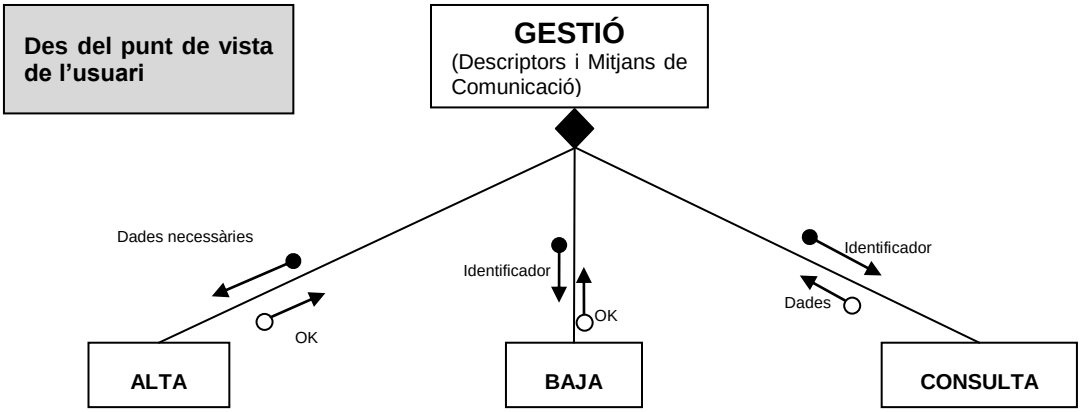

Figura 35: Diagrama d'estructura dels processos de Gestió des del punt de vista de l'usuari

Després d'observar el diagrama amb l'estructura general dels procediments de Gestió des del punt de vista de l'usuari (*Figura 35*), es passa a especificar els procediments que s'executen a cada mòdul per portar a terme les seves funcionalitats.

A l'igual que s'ha fet amb l'esquema general, el detall de cada esquema s'ha agrupat pels diferents subsistemes que s'han especificat anteriorment durant el transcurs de la documentació, ja que tal com s'ha comentat prèviament, els processos dels subsistemes de Gestió de Descriptors i de Mitjans de Comunicació són molt similars.

Mòdul d'Altes

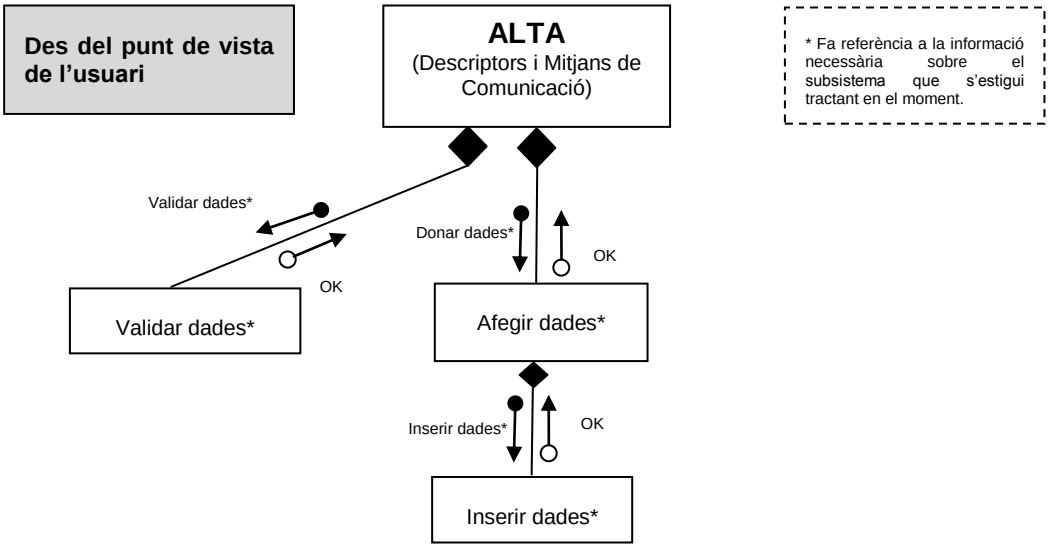

Figura 36: Diagrama d'estructura del mòdul d'altes (Usuari)

Associat a la *Figura 36* es troba el pseudocodi següent:

```
Procediment ALTA()
      dades←Obtenir_dades()
      Si (Validar_Dades(dades)) Llavors
            Afegir(dades)
            solució← "Alta correcte"
      Altrament
            solució← "Alta fallida"
      Fsi
      ← solució
Fprocediment
```

Aquest procediment d'alta el portarà a terme l'usuari quan vulgui afegir al sistema un descriptor o un mitjà de comunicació als seus seguiments de premsa. Així doncs el mètode *Obtenir_dades()* que es pot observar al pseudocodi recuperarà dels formularis vistos a les pantalles de les *Figures 28 i 29* (Pantalla de Gestió de Descriptors i Pantalla de Gestió de Mitjans de Comunicació), ja sigui afegint dades com la paraula que es buscarà al contingut de les notícies, com la selecció que es faci del llistat de mitjans de comunicació que es facilita.

Un cop el sistema ha recuperat la informació dels formularis corresponents, es farà una validació de que no s'ha deixat cap atribut necessari per la continuació del procediment, en cas afirmatiu es realitzarà la inserció a la Base de Dades de la nova informació i s'avisarà al formulari de que el procés s'ha finalitzat correctament. En el cas de que no es validin les dades, no es realitzarà la inserció i s'informarà pertinentment a l'usuari de que ha hagut problemes.

Mòdul de Baixes

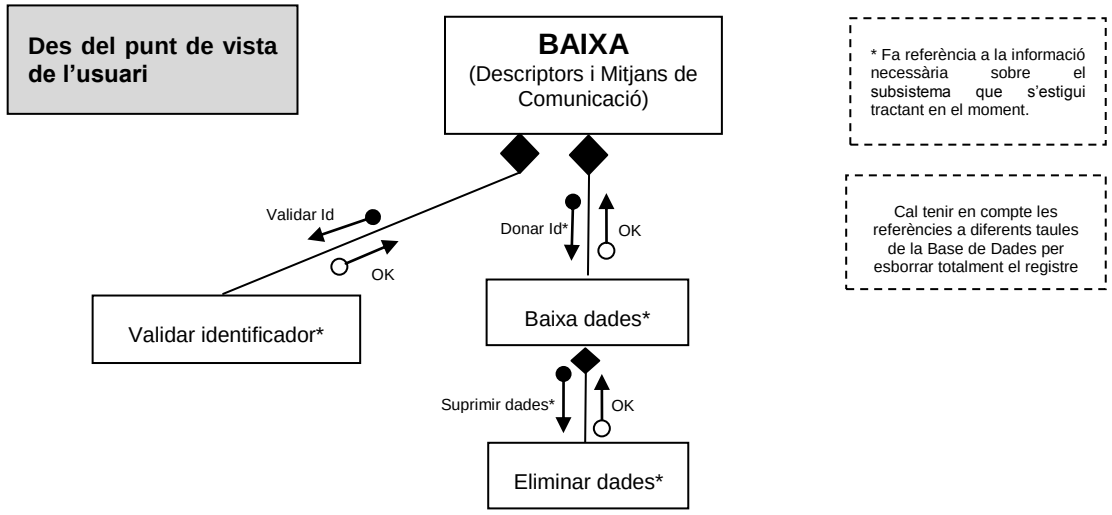

Figura 37: Diagrama d'estructura del mòdul de baixes (Usuari)

Aquesta *Figura 37* està recolzada amb el pseudocodi següent:

```
Procediment BAIXA()
      id←Obtenir_Identificador()
      Si (Validar_Identificador(id)) Llavors
            Eliminar(id)
            solució← "Baixa correcte"
      Altrament
            solució← "Baixa fallida"
      Fsi
      ← solució
Fprocediment
```

Aquest procediment que aquí es reflecteix és una mica més especial que l'anterior, ja que cal que tingui en compte un seguit de taules i la possible relació entre els seus registres.

De tal manera, un cop l'usuari ha escollit l'ítem o mitjà (del qual es vol donar de baixa en el seus sistema de seguiments de premsa) del llistat que l'ofereix la pantalla de Gestió de Descriptors o la de Gestió de Mitjans de Comunicació (*Figures 28 y 29*), cal que el sistema ho verifiqui com a pertanyent a la Base de Dades associada a l'aplicació i procedeixi a fer un recorregut en cascada per totes les taules de la Base de Dades a les que pot aparèixer el atribut seleccionat i decidir de quines taules s'esborra o no.

El criteri que es segueix per esborrar o no un registre d'una taula de la Base de Dades és el següent:

- Si l'atribut es troba associat a més usuaris o a alguna notícia, només s'esborrarà de la taula que l'associa amb l'usuari en qüestió que es vol donar de baixa.
- Si l'atribut no té cap associació amb cap altre usuari o notícia, s'esborrarà de la taula que el relaciona amb l'usuari i també de la seva taula pròpia. (Ja sigui DESCRIPTOR o MITJA)

Mòdul de Consultes

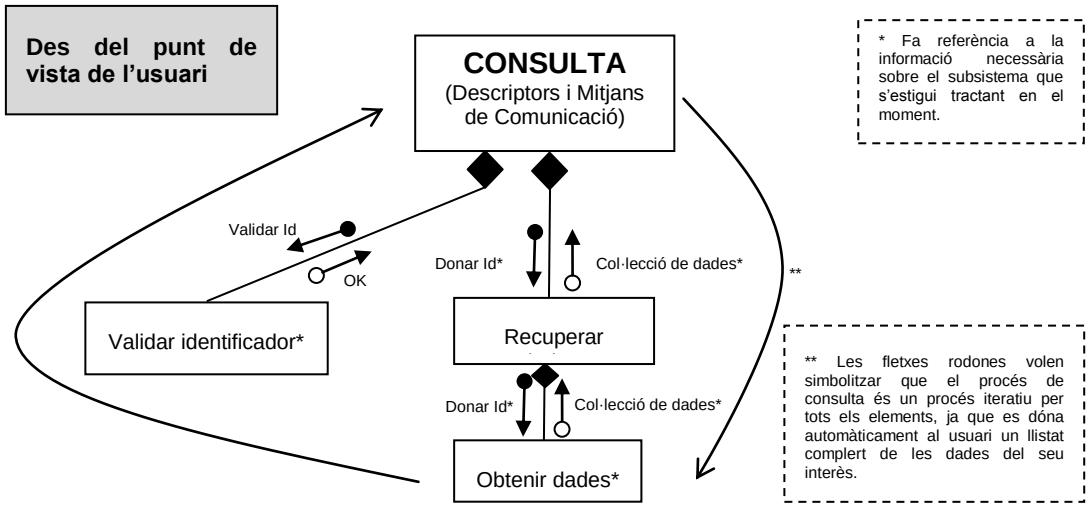

Figura 38: Diagrama d'estructura del mòdul de consulta (Usuari)

A continuació s'aportarà el pseudocodi que recolza al diagrama de la *Figura 38*:

```
Procediment CONSULTA()
      Per cada element fer
            id←Obtenir_Identificador()
            Si (Validar_Identificador(id)) Llavors
                  dades←ObtenirDades(id)
                  coleccio.afegir(dades)
            Fsi
      Fper
      ← coleccio
Fprocediment
```

Aquest pseudocodi reflecteix el procés de donar la informació a l'usuari sobre la col·lecció de dades que té incorporades al sistema (depenent del subsistema en el que es trobi). D'aquesta manera, el sistema recorrerà les taules amb la informació necessària per facilitar-la a l'usuari en forma de llistat a la pantalla corresponent de l'àrea privada del sistema.

Així doncs, aquest mòdul ofereix dos serveis alhora, el llistat de tota la informació que conté sobre l'àrea determinada on es troba l'usuari i al mateix temps el detall de cada registre que es té sobre ell. Això s'ha proporcionat així ja que les àrees de Gestió de Descriptors i Gestió de Mitjans de Comunicació tenen poques dades rellevants i es poden visualitzar d'una manera còmode per l'usuari, totes alhora i estructurades de forma clara i concisa.

Mòduls del Sistema des del punt de vista del propi Sistema

A continuació es presentarà el mòdul de gestió des del punt de vista del sistema (*Figura 39*). Aques mòdul conté processos que poden semblar iguals que els presentats des del punt de vista de l'usuari, però que a part d'afectar a una altre zona de l'aplicació (la Gestió de Notícies) també realitzen treballs sobre la Gestió de Mitjans de Comunicació que fa l'administrador del sistema per tenir la Base de Dades actualitzada i donar millor servei a l'usuari.

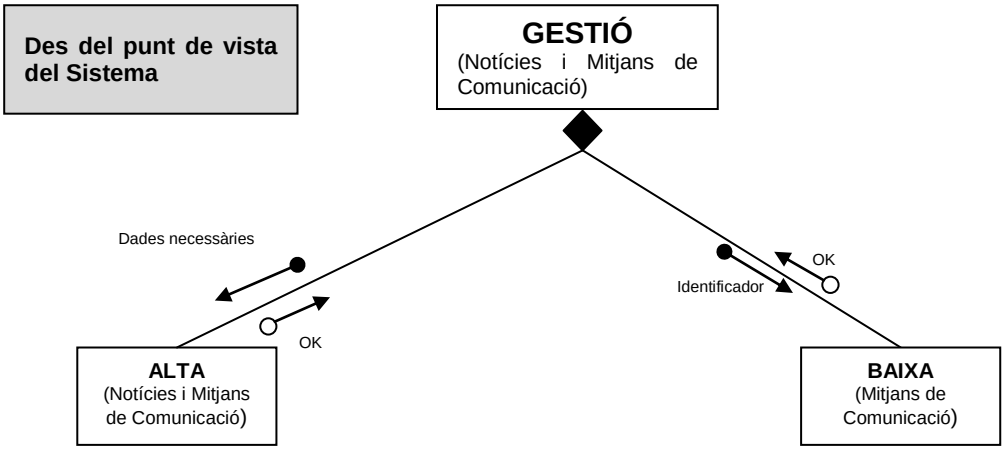

Figura 39: Diagrama d'estructura des del punt de vista del Sistema

De la mateixa manera que s'ha fet amb el punt de vista de l'usuari, es passa a analitzar mòdul a mòdul el diagrama d'estructura des del punt de vista del sistema. Des d'aquest punt de vista, les accions a realitzar són només d'alta i baixa a la Base de Dades, ja que la consulta i tot procés similar es fa des de la perspectiva de l'usuari.

A continuació es presentarà el detall dels mòduls del sistema que donen una visió clara dels processos més significatius del projecte, sobretot en el que respecta a donar d'alta al sistema un mitjà de comunicació o com donar d'alta una notícia, després de localitzar-la dins de la pàgina web a on es troba. Aquests processos porten tot el pes de l'aplicació, ja que són els que donen al projecte tot el seu valor afegit. Per tant, es passaran a comentar amb detall amb els diagrames d'estructura següents, juntament amb el se pseudocodi i explicacions associats.

Mòdul d'Altes

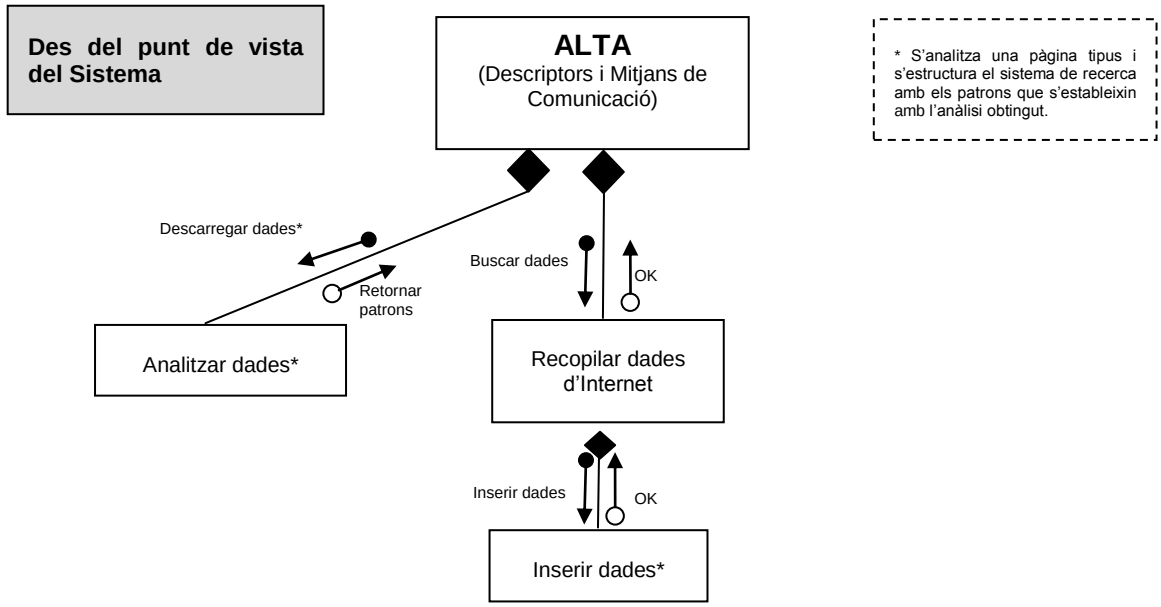

Figura 40: Diagrama d'estructura del mòdul d'altes (sistema)

En aquest cas, l'alta al sistema d'un mitjà de comunicació i una notícia seran recolzades per pseudocodis diferents. El que tenen en comú, es que prèviament, tal com s'observa a la *Figura 40*, s'haurà necessitat un anàlisi per configurar la programació amb els patrons que s'estableixin per poder realitzar la cerca, tant d'enllaços com d'estructura de la notícia.

Pseudocodi d'alta d'un mitjà de comunicació:

```
Procediment ALTA_MITJA()
      dades←observarWeb()  //Administrador del Sistema
      afegir(dades)
Fprocediment
```

En aquest primer pseudocodi es pot observar com l'administrador del sistema haurà de fer el procés d'observar la web del mitjà de comunicació que es vol afegir al sistema. Per poder tractar les dades de la pàgina web s'utilitzarà la llibreria cUrl [12](comentada a l'apartat de Coneixements Previs: 3.2.1) i els mètodes que facilita per descarregar el contingut i poder fer la selecció de la informació que interessa recopilar. Un cop es tenen les dades necessàries (Nom, descripció i direcció web d'on es troba) es dona d'alta al sistema el nou mitjà de comunicació fent el corresponent INSERT a la Base de Dades i ja es pot oferir a l'usuari com un mitjà més per als seus seguiments de premsa.

A continuació es presenta el pseudocodi corresponent a l'alta d'una notícia:

```
Procediment ALTA_NOTICIA()
      Descriptors←recuperarDescriptors()
      Per cada mitja.actiu de la BDD fer
            dades←descarregarCodiFontWeb()
            enllaços←recuperarEnllaçosNotícies(dades)
            Per cada enllaç fer
                  noticia←recuperarNoticia(enllaç)
                  Si apareix(descriptors) llavors
                        afegirNoticia(noticia,dadesGenerals)
                  altrament
                        afegirDescart(enllaç)
                  fsi
            fper
      fper
Fprocediment

Procediment recuperarEnllaçosNoticies(dades)
      Enllaços ← trobarTotesLesCoincidencies(patro,dades)
      ← enllaços
Fprocediment

Procediment afegirNoticia(noticia,dadesGenerals)
      titular ← trobarTotesLesCoincidencies(patro1,noticia)
      entradeta ← trobarTotesLesCoincidencies(patro2,noticia)
      cos ← trobarTotesLesCoincidencies(patro3,noticia)
      afegirALaBDD(titular,entradeta,cos,dadesGenerals)
Fprocediment
```

Amb el pseudocodi que aquí s'observa referent a l'alta d'una notícia, es pot veure els diferents passos que haurà de realitzar el sistema per poder arribar a obtenir una notícia.

Així doncs, la seqüència d'accions que s'hauran de portar a terme són les següents:

1. Es recuperen de la Base de Dades tots els descriptors pels quals s'haurà de filtrar les notícies.
2. Per cada mitjà actiu de la Base de Dades es descarrega el codi font per poder cercar els enllaços pertinents. Aquest enllaços seran trobats gràcies al procediment

recuperarEnllaçosNoticies que farà ús de mètodes propis del lleguatge PHP [13] com són la funció *preg_match_all*, representada al pseudocodi amb *trobarTotesLesCoincidencies*, que amb el patró que se li diu i un conjunt de dades retorna totes les coincidències que hi ha al conjunt de dades donat.

3. Per cada enllaç trobat, es descarrega el contingut i es filtra si s'hi troba algun dels descriptors. (un altre cop es filtra gràcies a la mateixa sentència de PHP)
    a. Si no hi és, només s'afegeix a la taula DESCART
    b. Si hi ha, s'afegeix la notícia al sistema amb el procediment que es pot observar a *afegirNoticia*. Primerament s'han de localitzar totes les parts importants que es volen emmagatzemar de la notícia. Això es portarà a terme mitjançant la funció ja comentada *trobarTotesLesCoincidencies* per cada part de la notícia que es vulgui recuperar (que vindrà localitzada per patrons diferents o no). Un cop es tenen recuperades les parts s'afegeix a la Base de Dades juntament amb les anomenades *dadesGenerals* que seran el mitjà origen de la notícia, la data de captació i els descriptors que conté. També, a la vegada que s'afegeix a la taula NOTICIA, s'afegeixen també totes les relacions pertinents amb CLIENT i DESCRIPTOR.

Un cop s'han seguit tots els passos per tots els enllaços de mitjans de comunicació de la Base de Dades es trobarà actualitzat el sistema per a que l'usuari pugui visualitzar una nova remesa de notícies.

Aquest procés del mòdul d'altes és un procés que l'administrador determina quant s'executa. Així doncs, cal un estudi del món dels portals per determinar el temps òptim de cada quan cal programar una nova cerca pel sistema.

Mòdul de Baixes

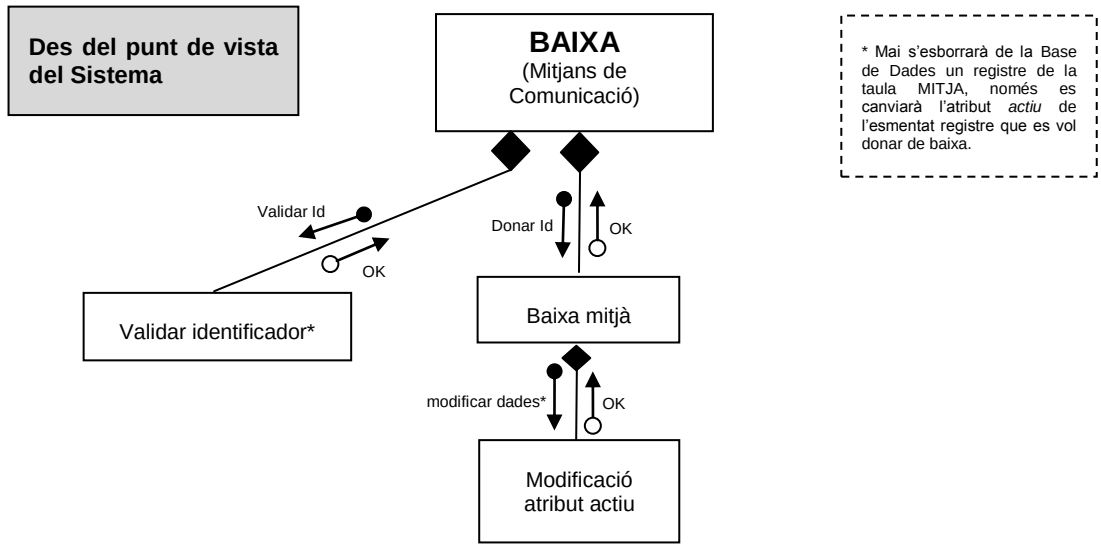

Figura 41: Diagrama d'estructura del mòdul de baixes (sistema)

A la *Figura 41* es pot observar el diagrama d'estructura del mòdul de baixes que realitza l'acció de donar de baixa un mitjà de comunicació del sistema, per iniciativa de l'administrador del mateix. Aquesta opció farà que cap usuari que no estigui donat d'alta en seguiments a aquest mitjà pugui fer-ho posteriorment al pas d'aquest procés.

A continuació es mostra el pseudocodi que recolza aquest diagrama:

```
Procediment BAIXA()
      id←Obtenir_Identificador()
      Eliminar(id)
Fprocediment
```

El procés de donar de baixa un mitjà de comunicació, s'ha determinat com un procés senzill per no deixar a nous usuaris donar-se d'alta en nous seguiments per aquest mitjà. Així doncs, l'administrador del sistema obtindrà el identificador del mitjà que vulgui donar de baixa i anirà al registre pertinent de la Base de Dades a canviar el seu atribut *actiu* per avisar als demés processos que utilitzen la taula MITJA de que aquell registre està momentàniament donat de baixa.

D'aquesta manera s'aconsegueix guardar la consistència de les dades pels processos que necessiten informació d'aquest mitjans de comunicació donats de baixa i dels quals, per exemple, ja es tenen notícies introduïdes al sistema.

7.8.2 Exemple de cerca d'enllaços

Després de veure com funcionen els diferents processos dels mòduls del Sistema es comenta amb detall el procediment d'obtenció d'un enllaç d'una pàgina web d'un mitjà de comunicació concret.

Tot seguit es visualitzarà la portada del mitjà i el seu codi font, també d'una notícia qualsevol.

Figura 42: Portada de LaVanguardia.es

L'administrador del sistema haurà d'anar al portada del mijtà de comunicació escollit i triar l'opció que li ofereix el menú: "Veure codi fon de la pàgina" (aquest menú contextual sortirà al prémer el botó dret de *mouse* sobre l'àrea de la pàgina, il·lustrat a la *Figura 42*).

Un cop s'obté el codi font, com es pot observar a la *Figura 43*

Figura 43: Codi Font de la portada de LaVanguardia.es

L'administrador del sistema haurà de configurar l'expressió regular necessària per obtenir el patró desitjat de cerca. Així doncs, amb l'anàlisi de la zona on es troben localitzats els enllaços que solen ser notícies s'arriba a la següent expressió:

```
"|a href=\"(http://www.lavanguardia.es/lv24h/(.*))\"|U"
```

A continuació s'analitza per parts el patró:

- `|a href=\ "` : determina com ha de començar la seqüència de caràcters per recaptar-la
- `(http://www.lavanguardia.es/lv24h/(.*))` : determina el conjunt de caràcters que es recopilaran. Aquest està dividit en dos conjunts determinats pels parèntesis, el conjunt que captura (.*) i que recupera qualsevol conjunt de caràcters i el conjunt complert amb la primera part determinat per (http://www.lavanguardia.es/lv24h/...)

Amb aquest patró i la funció de PHP preg_match_all, es recupera el següent del codi font de la pàgina web del mitjà de comunicació:

```
$array=preg_match_all($patro,$contingutWeb)
echo $array
```

Tenint present que la variable $contingutWeb té la informació descarregada de la pàgina web i la variable $patro conté l'expressió abans esmentada, el resultat de l'acció `echo $array` és:

```
Array→
[0][0]: a href="http://www.lavanguardia.es/lv24h/20070830/53389675955.html"
[0][1]: http://www.lavanguardia.es/lv24h/20070830/53389675955.html
[1][0]: a href="http://www.lavanguardia.es/lv24h/20070830/53389691751.html"
[1][1]: http://www.lavanguardia.es/lv24h/20070830/53389691751.html
[2][0]: a href="http://www.lavanguardia.es/lv24h/20070830/53389611032.html"
[2][1]: http://www.lavanguardia.es/lv24h/20070830/53389611032.html
    .
    .
    .
```

També es pot observar que el sistema per aconseguir els patrons per a les parts que formen la notícia. Ara es mostrarà l'estructura del codi font que té una notícia al mitjà de comunicació escollit per a obtenir els patrons.

Figura 44: Notícia de LaVanguardia.es

A continuació es mostra el codi font referent a la imatge de la *Figura 44*:

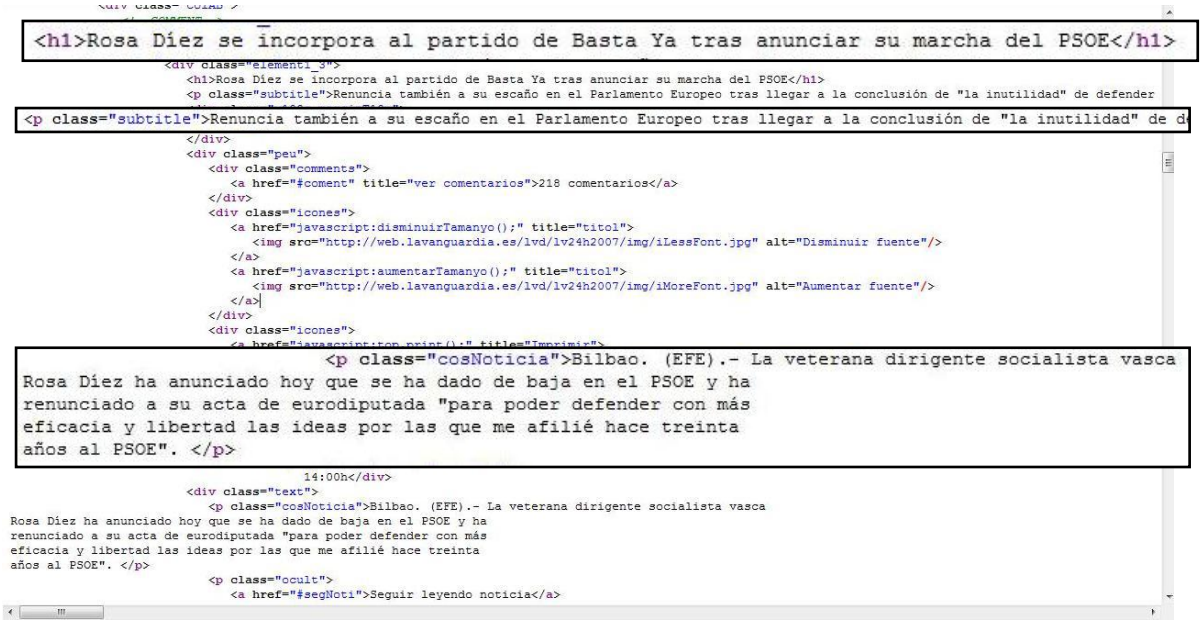

Figura 45: Codi Font de la notícia de LaVanguardia.es

Els patrons obtinguts per a localitzar cada part de la noticia són:

- Patró del titular: `"|element1_3\">\<h1\>(.*)\<\\h1|U"`
- Patró del entradeta: `"|class=\"subtitle\"\>(.*)\<\\p|U"`
- Patró del cos: `"|cosNoticia\">(.*)\<\\p|U"`

I al igual que succeeix amb el cas dels enllaços, un cop s'apliqui el patró, aquest retornarà una matriu amb tantes cel·les com grups s'hagin posat a l'expressió (delimitats pels parèntesis).

En aquest cas que es mostra a la *Figura 45*, cada patró, al ser aplicat amb la funció corresponent de PHP (*preg_match_all*), recuperaria el següent:

Titular:

"Rosa Díez se incorpora al partido de Basta Ya tras anunciar su marcha del PSOE"

Entradeta:

"Renuncia también a su escaño en el Parlamento Europeo tras llegar a la conclusión de "la inutilidad" de defender sus ideas dentro del partido"

Cos:

"Bilbao. (EFE).- La veterana dirigente socialista vasca
Rosa Díez ha anunciado hoy que se ha dado de baja en el PSOE y ha
renunciado a su acta de eurodiputada "para poder defender con más
eficacia y libertad las ideas por las que me afilié hace treinta
años al PSOE"."

"En una conferencia de prensa en Bilbao, Díez destacó que ha
llegado "a la conclusión de la inutilidad de defender estas ideas
dentro del partido" y precisó que si hay que elegir entre "seguir
las instrucciones del partido o cumplir el compromiso adquirido con
los ciudadanos", se queda con lo segundo".

 … "[23]

[23] La noticia continua i la recuperaria tota el patró, en aquest cas del cos, hi ha dues zones que la formen part.

7.9 Disseny físic de dades (v.1)

7.9.1 Disseny del Model Físic de Dades (v.1)

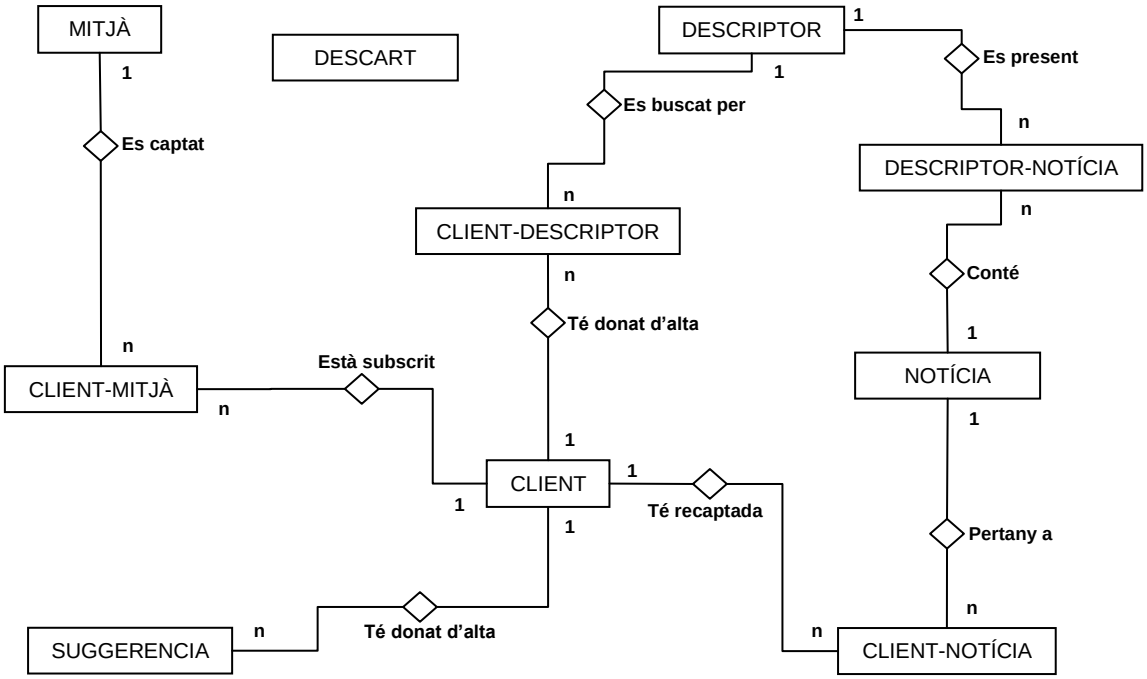

Figura 46: Diagrama d'Entitat – Relació

7.9.2 Descripció de les Taules (v.1)

Mitja

Camp	Tipus	Descripció
MITJA_ID	Smallint (5)	Identificador de la taula, està configurat per a que tingui autoincrement
MITJA_NOM	Varchar (100)	Nom del mitjà de comunicació
MITJA_DESCRIPCIO	Varchar (200)	Descripció del mitjà de comunicació
MITJA_URL	Varchar (150)	Enllaç a on es localitza el mitjà de comunicació
MITJA_ACTIU	Tinyint(1)	Número que indica si el mitjà està actiu per realitzar les cerques

Taula 14: Mitja (I)

Aquesta taula emmagatzema tota la informació necessària al respecte d'un mitjà de comunicació. D'aquesta manera, en aquesta versió, només són necessàries les seves dades principals com són el nom, la descripció i la direcció a on es troba localitzat el mitjà a Internet. També, cal especificar que el camp MITJA_ACTIU és un camp informatiu per a la programació que diu al sistema si aquest mitjà està habilitat per a la cerca o no.

Client

Camp	Tipus	Descripció
CLIENT_ID	Smallint (5)	Identificador de la taula, està configurat per a que tingui autoincrement
CLIENT_NOM	Varchar (150)	Nom complet del client (Nom i cognoms)
CLIENT_DIRECCIO	Varchar (300)	Direcció física d'on resideix el client
CLIENT_TELEFON	Smallint (6)	Telèfon de contacte del client
CLIENT_FAX	Smallint (6)	Fax del client
CLIENT_MAIL	Varchar (150)	Direcció de correu electrònic del client
CLIENT_ACTIU	Tinyint(1)	Número que indica si el client està actiu per realitzar-li les cerques
CLIENT_USER	Varchar(100)	Nom d'usuari que s'ha posat el client
CLIENT_PASS	Varchar(100)	Contrasenya d'usuari que s'ha posat el client xifrada
CLIENT_RECIVE	Varchar(2)	Determina si el client vol rebre informació al correu electrònic

Taula 15: Client (I)

La taula que aquí es visualitza al final de la pàgina anterior (*Taula 15*) representa l'esquema que es necessita per guardar a la Base de Dades tota la informació referent a l'usuari o client del sistema. Dades generals com poden ser el nom, la direcció, el telèfon, el fax, ... i dades més específiques per a que l'usuari pugui utilitzar el portal i moure's per la seva àrea de client, com són el CLIENT_USER i CLIENT_PASS. També es troben dos atributs especials: CLIENT_RECIVE que diu al sistema si l'usuari vol rebre informació al seu correu electrònic i CLIENT_ACTIU que li diu al sistema si l'usuari en qüestió està habilitat per cercar-li les notícies que s'ajusten als seus descriptors i mitjans de comunicació associats i introduïts a les taules corresponents.

Noticia

Camp	Tipus	Descripció
NOTICIA_ID	Smallint (5)	Identificador de la taula, està configurat per a que tingui autoincrement.
NOTICIA _TITULAR	Varchar (200)	Contingut del titular de la notícia
NOTICIA _ENTRADETA	Varchar (300)	Contingut de l'entradeta de la notícia
NOTICIA _COS	Text	Contingut del cos de la notícia
NOTICIA _URL	Varchar (150)	Direcció electrònica d'on s'ha trobat la notícia
NOTICIA _DATA	Date	Data a la que s'ha captat la notícia d'Internet

Taula 16: Noticia (I)

A la versió 1 del sistema es troba aquesta taula NOTICIA a on es poden veure els seus atributs. Els que es recullen d'Internet són el titular, l'entradeta i el cos de la notícia, també es recapta l'enllaç a on s'ha trobat. Per últim s'emmagatzema la data a la qual s'ha emmagatzemat la notícia al sistema.

Descriptor

Camp	Tipus	Descripció
DESCRIPTOR_ID	Smallint (5)	Identificador de la taula, té autoincrement.
DESCRIPTOR _NOM	Varchar (100)	Paraula o paraules que s'utilitzaran per la cerca de notícies
DESCRIPTOR _DESCRIPCIO	Varchar (300)	Descripció del camp anterior

Taula 17: Descriptor (I)

La taula Descriptor és l'encarregada de guardar tot el referent a les paraules que els usuaris volen cercar per la xarxa. Així doncs, és necessari guardar la paraula o paraules a cercar i una breu descripció que l'usuari vulgui aportar a aquestes.

Suggeriment

Camp	Tipus	Descripció
SUGGERIMENT_ID	Bigint (20)	Identificador de la taula, té autoincrement.
SUGGERIMENT _COS	Text	Contingut del cos del suggeriment
SUGGERIMENT _ID_CLIENT	Smallint (5)	Identificador de la taula CLIENT que fa referència a qui l'ha fet

Taula 18: Suggeriment (I)

Aquesta taula recull els suggeriments dels usuaris del sistema, de tal manera, emmagatzema els comentaris que des de l'aplicació els usuaris envien mitjançant l'àrea corresponent del sistema.

Descart

Camp	Tipus	Descripció
DESCART_ID	Bigint (20)	Identificador de la taula, té autoincrement.
DESCART _URL	Varchar (150)	Direcció electrònica de la notícia descartada pel sistema.

Taula 19: Descart

Aquesta taula emmagatzemarà els enllaços trobats pel sistema durant la seva cerca i que han resultat infructuosos, és a dir, que el sistema no ha trobat cap coincidència entre el contingut de la notícia i els descriptors de la Base de Dades.

A partir de la taula que es presenta a continuació, es reflecteixen les relacions n-m que hi ha al diagrama de classes dibuixat a la *Figura 46*.

Client-Mitja

Camp	Tipus	Descripció
CLIENT_MITJA_ID_CLIENT	Smallint (5)	Identificador de la taula CLIENT
CLIENT_MITJA_ID_MITJA	Smallint (5)	Identificador de la taula MITJA

Taula 20: Client_Mitja (I)

Client-Descriptor

Camp	Tipus	Descripció
CLIENT_DESCRIPTOR_ID_CLIENT	Smallint (5)	Identificador de la taula CLIENT
CLIENT_DESCRIPTOR_ID_DESCRIPTOR	Smallint (5)	Identificador de la taula DESCRIPTOR

Taula 21: Client Descriptor (I)

Mitja-Noticia

Camp	Tipus	Descripció
MITJA_NOTCIA_ID_MITJA	Smallint (5)	Identificador de la taula MITJA
MITJA_NOTICIA_ID_NOTICIA	Smallint (5)	Identificador de la taula NOTICIA

Taula 22: Mitja-Noticia (I)

Descriptor-Noticia

Camp	Tipus	Descripció
DESCRIPTOR_NOTÍCIA_ID_DESCRIPTOR	Smallint (5)	Identificador de la taula DESCRIPTOR
DESCRIPTOR_NOTÍCIA_ID_NOTICIA	Smallint (5)	Identificador de la taula NOTICIA

Taula 23: Descriptor-Noticia (I)

8 Construcció del Sistema d'Informació (v.1)

8.1 Execució de les proves unitàries (v.1)

Tal com l'enunciat d'aquest punt india, les probes unitàries són aquelles que es realitzen punt per punt dels processos que es troben al sistema. D'aquesta manera es poden arribar a detectar els errors al nivell més baix possible. També facilita tot el que esdevé l'aïllament de qualsevol problema i la correcció d'aquest de manera eficaç.

Per portar a terme una bona execució de les proves unitàries, per analitzar el projecte, cal seguir un protocol d'accions seguint un procés lineal i tenint com a principal objectiu les àrees a on es modifica la informació de la Base de Dades. D'aquesta manera es declara un protocol a seguir per fer una comprovació de tots els aspectes del programa:

- Accés a les diferents àrees:
    - Sinó es realitza cap acció, el sistema es manté a la pàgina actual i no accedeix a cap àrea nova.
    - Si es selecciona un enllaç, el sistema ha de redirigir-se cap a l'àrea desitjada per l'usuari.
- Als processos que l'usuari es dóna d'alta tots els paràmetres necessaris estan omplerts:
    - Sinó, mostrar el missatge corresponent i no realitza cap acció.
    - Si hi és tot correcte, executar l'acció de donar d'alta i mostrar el missatge pertinent.
- Als processos de baixa cal controlar que:
    - Comprovar que s'ha escollit algun registre a eliminar.
    - Una vegada seleccionat el registre i executada l'acció, mostrar a l'usuari el missatge informatiu corresponent.
- Al processos de consulta, respectar que:
    - Quan l'usuari entra a l'àrea pertinent troba el llistat de consulta a punt, amb totes les opcions de tractament que ofereix el sistema actives (donar de baixa en tots els casos menys en la notícia, que serà poder consultar).

8.2 Execució de les proves d'integració (v.1)

Una vegada s'han portat a terme les proves unitàries i solucionats tots els problemes sorgits d'aquestes, es passaran a realitzar les proves d'integració. Aquestes proves consisteixen en verificar que tots els processos de l'aplicació realitzin les accions correctament. Per portar a terme aquestes proves es passa a comparar el resultat de les accions amb el contingut que es tindria a una Base de Dades replicada, fent la mateixa acció directament sobre ella.

Per poder fer la comparativa en aquest dos medis diferents cal tenir en compte diferents aspectes:

- El sistema muntat en aquest projecte està en un entorn web controlat programat per donar un servei fàcil a l'usuari.
- Les proves a la Base de Dades s'han de realitzar a base de consultes i accions directes sobre ella i per no alterar el contingut de la mateixa serà necessari fer les proves sobre una rèplica.

- Per a que les proves es puguin donar per bones cal que la crida que s'executa a la Base de Dades sigui la mateixa que es fa al sistema implementat al projecte, això vol dir que cal que el sistema estigui preparat per retornar aquesta crida a l'administrador per a que pugui realitzar les proves.

Un cop es tenen complits els requisits de les proves d'integració, es pot passar a la posta en marxa del sistema. En aquest cas, donat el caire del projecte, després de les proves d'integració i d'haver analitzat el sistema muntat, s'ha vist com es podia millorar el sistema, ja que la forma de treball descrita deixava de ser eficaç al intentar analitzar un mitjà de comunicació que no fos l'original de l'anàlisi (Fase 5 del Diagrama de Gantt, *Taula 2* i *Figura 3* de la secció 5.1.2 Definició del Pla de Treball)

Per aquesta raó s'ha decidit fer una millora en el sistema que es passarà a comentar a continuació.

Segona Versió del desenvolupament

9 Anàlisi i Disseny del Sistema d'Informació (v.2)

9.1 Definició del Sistema (v.2)

9.1.1 Determinació de l'Abast del Sistema (v.2)

Un cop s'arriba a aquest punt es veu que en la versió inicialment desenvolupada es troben errors en el funcionament. Aquests errors apareixen al intentar realitzar la recaptació de notícies d'un altre mitjà que no sigui l'original amb el que s'ha començat l'anàlisi.

Era una possibilitat que succeís un problema d'aquest caire, ja que la gent que treballa amb l'HTML no segueix estàndards i el codi font de les pàgines web no segueix la mateixa estructura ni jerarquia, ja que es tenen moltes possibilitats de programar una pàgina web i que el resultat final (visualment) sigui el mateix.

Així doncs, l'abast del sistema que es vol obtenir, continua sent el mateix, i el que canviarà serà la forma d'arribar a obtenir-ho. Igualment que passa amb la primera versió del desenvolupament, els desitjos de l'usuari final continuen essent:

- *Obtenir notícies*: aquest és el fi bàsic del projecte, el motiu pel qual es desenvolupa i el que principalment ha desencadenat aquesta segona versió que es passa a documentar.
- *Gestionar els mitjans de comunicació*: per poder portar una gestió i ordre de les notícies que obté cada usuari, cal facilitar-li la gestió dels mitjans de comunicació que té al seu abast per fer els seguiments de premsa que desitgi, aquest requisit també patirà canvis en els seus processos.
- *Gestionar els descriptors:* per últim es torna a trobar aquesta gestió, que és una conseqüència lògica del primer punt, i que serà l'únic requisit que no patirà canvis en la seva estructura de processos en aquesta segona versió del sistema.

Tot seguit es mostra el nou diagrama de context (*Figura 47*) on es pot començar a intuir els canvis en els processos d'obtenció de notícies i de gestió dels Mitjans de Comunicació

Diagrama de Context del Sistema (v.2)

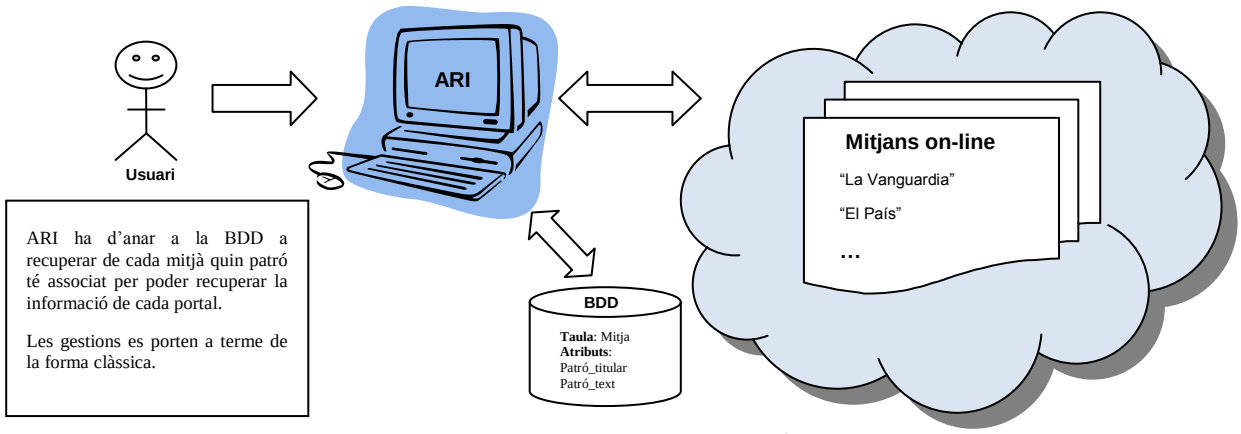

Figura 47: Diagrama de Context del Sistema (Versió 2)

9.1.2 Identificació d'Usuaris Participants i Finals (v.2)

Els usuaris participants en el procés de disseny en aquesta segona versió del desenvolupament, continuen sent els mateixos que a la primera versió:

- El responsable de l'empresa que va iniciar el projecte i que era el que va determinar quines eren les prioritats d'implementació i d'obtenció de mitjans de comunicació.
- Jo mateixa, com a desenvolupadora del projecte i co-responsable.

9.2 Establiment de Requisits (v.2)

En aquesta revisió dels requisits que ha d'afrontar el sistema, s'inclou un nou obstacle que no s'havia tingut en compte a la primera versió del desenvolupament: la NO homogeneïtat entre pàgines web. En un principi, el desenvolupament s'ha encarat a la recaptació seguint uns paràmetres estàtics establerts a la programació, ja que es creia que al estar analitzant un llenguatge d'etiquetes en webs dedicades al mateix totes farien servir les mateixes i, per tan, els mateixos patrons de cerca serien bons per les pàgines web d'uns mitjans de comunicació com per altres.

Així doncs, els requisits generals de l'aplicació continuen essent els mateixos però amb la necessitat de revisió dels dos últims punts:

- La transparència d'accés i gestió per a tenir organitzat tot el referent als mitjans de comunicació als quals s'està accedint durant el procés de recerca, el mateix al que ítems o descriptors es refereix, han d'estar introduïts a la Base de Dades.
- Una **Base de Dades que recolzi tot el procés** d'una manera eficaç i rendible, que faci de tot el procés de gestió quelcom natural. Pel que fa a l'obtenció de notícies, cal una estructura que afronti les dificultats de tenir administrat un gran volum d'informació, tota relacionada entre sí.
- Una **algorísmica eficient**, que faci que la cerca per la Xarxa sigui el més senzill i ràpid possible.

Un cop s'han establert de nou els requisits bàsics del sistema es passarà a fer una re-especificació dels processos que han sofert canvis per poder solucionar els problemes trobats a la primera versió.

9.2.1 La Gestió dels Mitjans de comunicació (v.2)

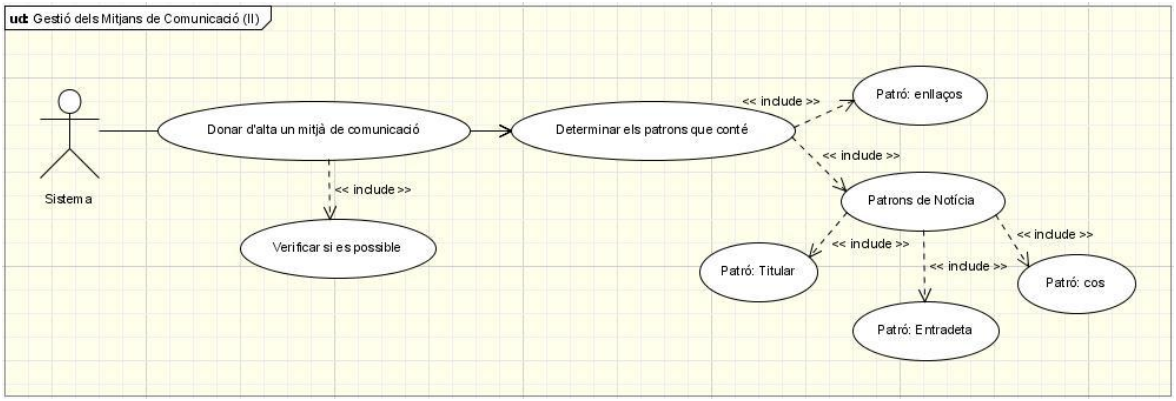

Figura 48: Cas d'ús – Gestió de Mitjans de Comunicació (II)

A aquesta nova versió del Cas d'ús: Gestió de Mitjans de Comunicació (*Figura 48*), només s'observa la part que afecta al sistema, i que es la que canvia. La part de la gestió que afecta a l'usuari continua sent igual que la descrita a la primera versió.

També cal dir que si s'observa el diagrama, el cas d'ús referent a donar de baixa un mitjà de comunicació, també continua sent igual que el descrit a la primera versió del desenvolupament i que només es descriu la nova versió del cas d'ús d'alta d'un mitjà de comunicació.

Cas d'ús	Alta d'un mitjà de comunicació (II)
Versió	1
Descripció	Donar d'alta un mitjà de comunicació.
Actors	Administrador del sistemaSistema.
Precondició	El sistema està connectat a la xarxa.
Flux Principal	1. Descarregar el codi font de la pàgina principal del mitjà de comunicació 2. Analitzar el codi de la pàgina principal. a. Estructura dels enllaços. b. Confecció del patró per localitzar-los. 3. Descarregar el codi d'una notícia del mitjà de comunicació. a. Estructura d'una notícia. b. Confeccionar els patrons de les parts de la notícia. 4. **Guardar els patrons al corresponent registre del mitjà de comunicació nou.** 5. **Estructurar el codi del sistema per a que agafi cada configuració de patrons.**
Postcondició	Codi confeccionat de manera que cerqui mitjançant els patrons.

Taula 24: Fitxa – Alta d'un Mitjà de Comunicació (II)

A la *Taula 24* es pot observar com el flux principal del cas d'ús ha patit un canvi respecte a la versió principal. Això ve donat perquè les pàgines web dels mitjans de comunicació no són homogènies. Per tant, cal guardar a la Base de Dades d'alguna manera quins patrons afecten a cada mitjà analitzat. I per tant, caldrà que es modifiqui tant la Base de Dades com el codi relacionat amb aquest procés.

9.2.2 L'obtenció de Notícies (v.2)

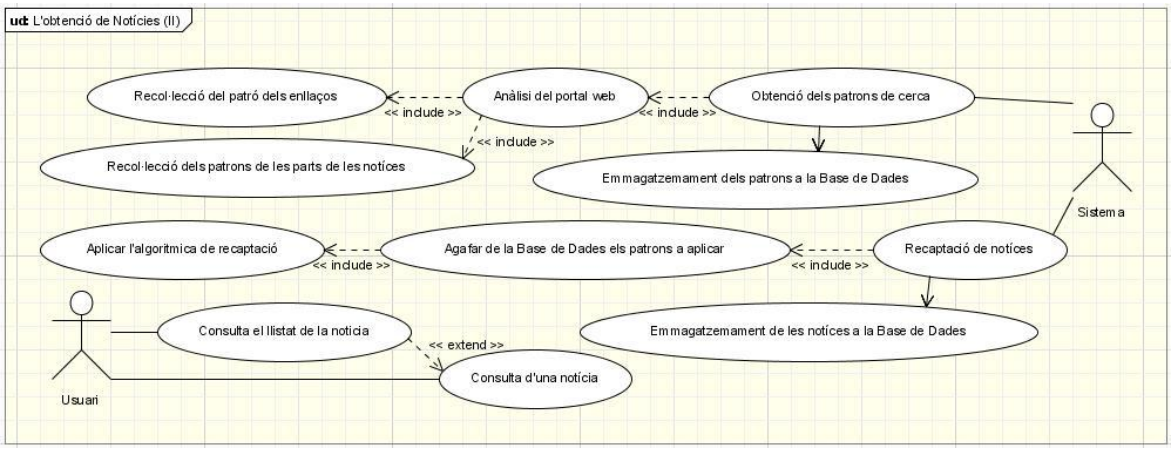

Figura 49: Cas d'ús – L'obtenció de Notícies (II)

Aquest cas d'ús reflectit a la *Figura 49* mostra els canvis patits per part del sistema en el que respecta a l'obtenció de notícies de la Xarxa. Tot seguit es mostra la fitxa corresponent a aquesta àrea del cas d'ús, a on queden clars els canvis que s'han portat a terme.

Cas d'ús	Recaptació de notícies (II)
Versió	1
Descripció	Recol·lectar les notícies de les webs dels mitjans de comunicació.
Actors	Sistema.
Precondició	El sistema està connectat a la xarxa.
Flux Principal	1. **Recuperar de la Base de Dades els patrons referents al mitjà de comunicació que es vol extreure la informació.** 2. Descarregar el codi font de la pàgina principal del mitjà de comunicació 3. Extreure els enllaços corresponent a notícies **mitjançant el patró d'enllaços que s'ha recuperat.** 4. Descarregar el codi d'una notícia del mitjà de comunicació. 5. Filtrar si s'ha d'emmagatzemar (segons apareixen els descriptors, o no) 6. Si apareixen els descriptors, extreure la informació necessària de la notícia **mitjançant els demés patrons extrets de la Base de Dades.** a. Extreure el titular b. Extreure l'entradeta c. Extreure el text 7. Emmagatzemar al sistema amb les relacions pertinents.
Postcondició	Descriptor esborrat del sistema

Taula 25: Fitxa – Recaptació de notícies (II)

Aquest procés de recaptació ha sofert un canvi en la seva algorítmica. Aquest ha vingut donat per la modificació de la Base de Dades, canvis en l'estructura de la taula MITJA.

Ara, el sistema ha de remetre's a la taula MITJA per poder recuperar del registre el patró que s'ajusta a les necessitats del mitjà de comunicació del qual s'està recaptant les notícies en comptes de tenir un patró predefinit per a tots els mitjans com succeïa a la primera versió del desenvolupament.

9.3 Identificació de Subsistemes d'Anàlisi (v.2)

9.3.1 Identificació i Definició de Subsistemes (v.2)

Tot seguit es passa a identificar les noves característiques que formen part dels subsistemes ja comentats a la primera versió del desenvolupament.

Els subsistemes que formen part del projecte continuen sent els mateixos, sobretot en el que es refereix al subsistema de Descriptors. Pel que fa als subsistemes de Mitjans de Comunicació i Notícies sofriran petits canvis en el desenvolupament d'algun dels seus processos.

A continuació es pot observar el diagrama de Flux de Dades a nivell general del sistema (*Figura 50*)

Diagrama de Flux de Dades de nivell 1 (v.2)

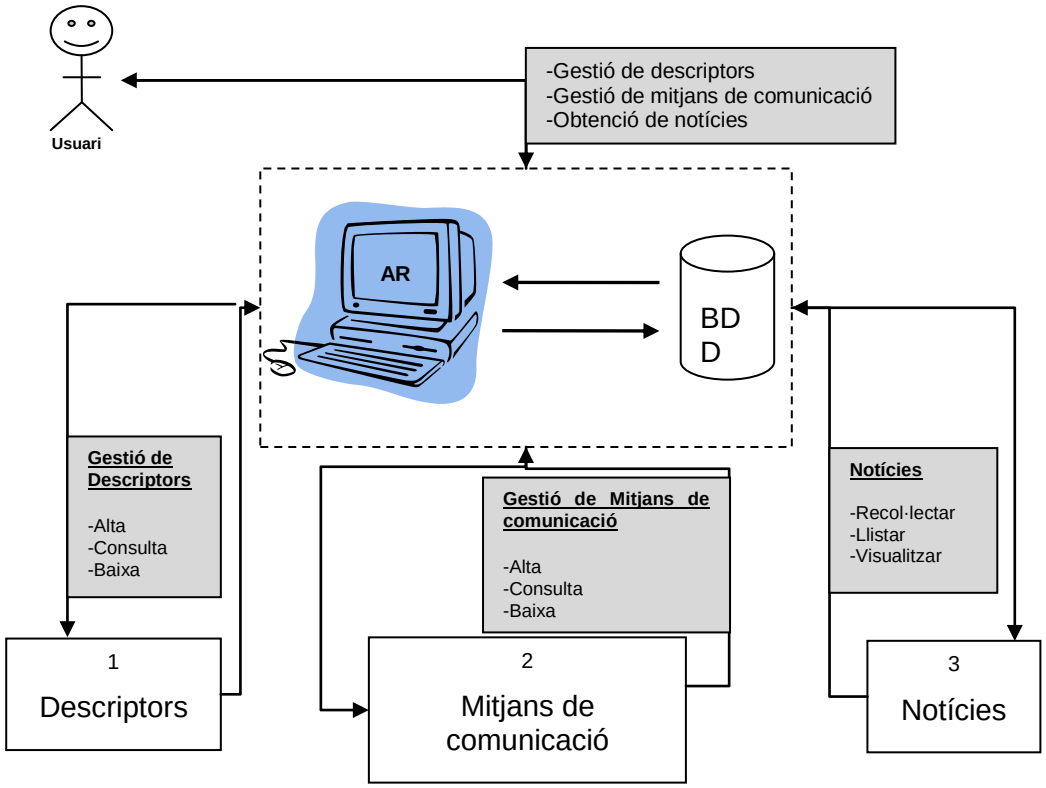

Figura 50: Diagrama de Flux de Dades de Nivell 1 (v.2)

Com es pot observar, aquest diagrama és igual al diagrama de Flux de Dades de la primera versió. Això ve donat perquè el que canvia no és el traspàs d'informació dins del sistema, ni les dades que finalment arriben a l'usuari, sinó que són els processos els que canvien la seva forma de treballar dins del sistema. Així doncs, la descripció de cada subsistema concorda amb l'exposada al punt 7.3.1 d'aquesta documentació.

9.4 Elaboració del Model de Dades (v.2)

9.4.1 Model Lògic de Dades Demanat (v.2)

Un cop es té clar que els subsistemes continuen essent els mateixos i que seran els procediments que executen el que canvia, es passa a examinar el model lògic de dades establert a la primera versió.

Amb aquest punt passa el mateix que amb la identificació dels subsistemes, en aparença tot segueix igual, el diagrama de classes d'aquesta versió continua sent (a nivell de classes) igual que el de la primera versió, tal com es pot observar a la *Figura 51* següent:

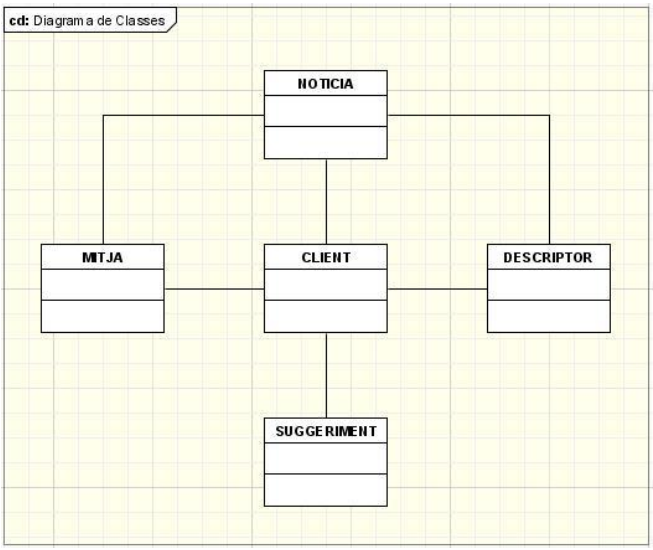

Figura 51: Diagrama de Classes (II)

El diagrama de classes, tal com succeeix a la primera versió, està format per poques classes i les que porten tot el pes fonamental són dues, NOTICIA i CLIENT, amb aquestes dues es distribueixen les accions bàsiques per dur a terme les necessitats de l'usuari de l'aplicació.

Encara que el diagrama continua sent igual que a la primera versió, els mètodes que conté cada classe canvien. Això passa per donar al sistema noves funcionalitats i poder realitzar els procediments d'una nova més eficaç a la de la primera versió del desenvolupament.

Amb això es vol aconseguir que el sistema pugui recuperar de la Xarxa tota la informació. Així, en la secció que es presenta a continuació s'observarà el detall dels processos que canvien, com ho fan i per què. També es veurà el detall de les taules que canvien a la Base de Dades i que fan que el sistema pateixi aquesta evolució que es presenta a la nova versió.

9.5 Elaboració del Model de Processos (v.2)

En aquest punt de la segona versió del desenvolupament es passa a explicar l'essència dels canvis que ha sofert el sistema, ja que a continuació s'explicarà el model de processos que engloba tota l'aplicació, centrant l'atenció en els que s'han vist afectats pels canvis en l'algorítmica.

Per tenir una visió més específica de les diferents àrees del model de processos s'ha dividit el model en els diferents subsistemes, per fer una explicació més acurada i clarificadora. Cal esmentar que el model de processos referent a la Gestió de Descriptors no canvia d'una versió a l'altre i que per tant ara no es fa esment.

9.5.1 Gestió dels Mitjans de Comunicació (v.2)

L'objectiu d'aquest subsistema continua sent el mateix que a la versió anterior i es portarà a terme amb el mateix tipus de processos. Fins i tot, els processos que es realitzen des del punt de vista de l'usuari són exactament iguals en tots els aspectes. Els que canvien són els referents al sistema, sobretot el procés de donar d'alta un nou mitjà de comunicació al sistema, a on l'algorítmica que hi ha darrera canvia substancialment.

A continuació es mostra el detall d'aquest procés que es troba dins d'aquest subsistema:

Procés 1: donar d'alta un Mitjà de Comunicació.

En aquesta versió es troba que el sistema ha de donar d'alta el mitjà de comunicació a la Base de Dades per a que l'aplicació pugui fer la cerca. Aquest procés el segueix tinent que dur a terme l'administrador del sistema que haurà d'incloure a la Base de Dades la informació necessària del nou mitjà de comunicació que es vulgui donar d'alta: nom, descripció, direcció web o enllaç i si està actiu o no. La diferència radica que ara cal un anàlisi més exhaustiu de la seva morfologia HTML, com està estructurat el seu codi font a la web i extreure del seu anàlisi un seguit de patrons:

- Patró identificador d'enllaços de notícies
- Patró identificador de titulars de notícies.
- Patró identificador d'entradetes de notícies.
- Patró identificador del cos de les notícies.

Tot seguit, a la *Figura 52* es mostra el global de processos que engloben aquest subsistema i que només ha patit canvis el primer.

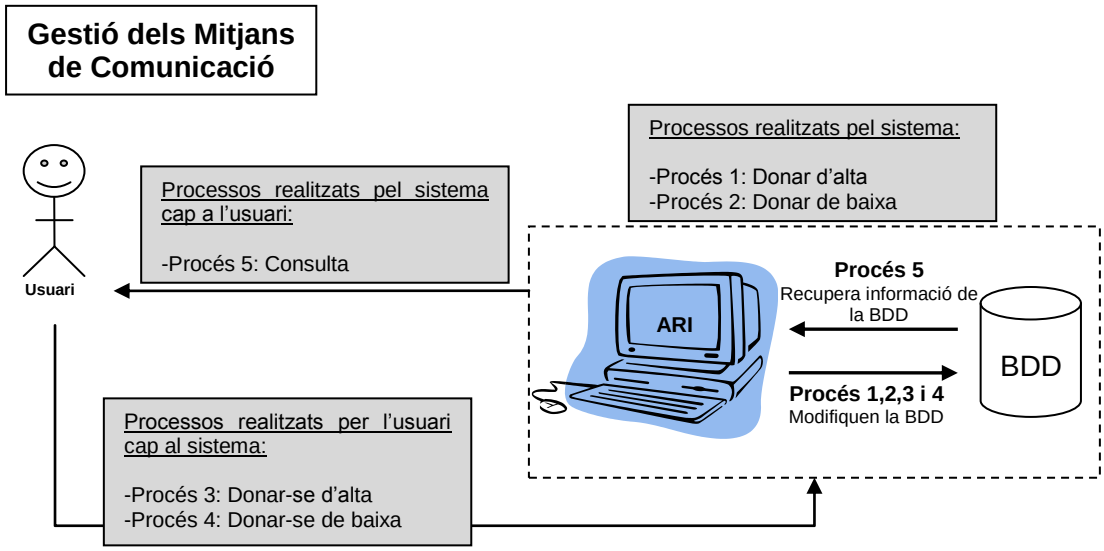

Figura 52: Digrama de Flux de Dades de Nivell 2 – Gestió de Mitjans de Comunicació (II)

9.5.2 Gestió de Notícies

Aquest subsistema, al igual que succeeix amb la primera versió, es diferencia dels altres dos subsistemes en que els seus processos són els encarregats d'obtenir la informació d'Internet. En aquesta segona versió del desenvolupament, aquest procés d'anar a la Xarxa a recuperar les notícies ha canviat substancialment en el que a l'algorítmica respecta.

En essència, dels processos que s'han comentat en la primera versió tornen a repetir-se en aquesta nova versió. Aquest es troben reflectits al diagrama de Flux de Nivell 2 (*Figura 53*) i es passa a comentar en detall el procés de recopilar notícies que és el que ha canviat la seva manera de treballar.

Procés 1: recopilar Notícies d'Internet

Aquest procés és el més important de tot el projecte i és el que ha provocat aquesta iteració en el desenvolupament del mateix. Aquest procés està comandat per el sistema i és l'encarregat d'anar-se per les pàgines web dels mitjans de comunicació escollits i buscar les notícies que corresponguin als descriptors que es desitja trobar. Tot això es podrà veure amb detall amb el pseudocodi aportat al punt 9.8.1 (Elaboració del Model de Processos)

En un inici, a la primera versió, es va pensar que les pàgines serien homogènies entre un mitjà de comunicació i un altre i es va decidir una manera de procedir que va donar un bon resultat per la recerca de notícies del mitjà de comunicació que es va agafar d'exemple per recuperar les primeres notícies d'Internet. Però al intentar aplicar el mateix algorisme per la recaptació d'un altre mitjà de comunicació es va veure com no funcionava, la qual cosa va provocar l'anàlisi d'altres mitjans de comunicació per poder arribar a la conclusió de que cada pàgina web de cada mitjà de comunicació pot tenir una estructura de codi font diferent.

Això ha provocat canvis a diversos nivells, per començar es procedeix a comentar les passes que en aquesta nova versió el sistema ha de portar a terme per recuperar les notícies:

1. *Recuperar els patrons de la Base de Dades associats a cada Mitjà de Comunicació*: cal que el sistema primerament vagi a la Base de Dades a recuperar la informació sobre els patrons que estan associats al codi font del mitjà de comunicació del qual es vol recuperar la informació
2. *Recuperar enllaços dels Mitjans de Comunicació*: cal qui el sistema vagi a la Base de Dades i recuperi la direcció web, l'enllaç, a on es pot trobar el contingut de cada mitjà.
3. *Descarregar el contingut de l'enllaç*: cal que el sistema descarregui la informació de la pàgina web i trobi tots els enllaços, aplicar-li el patró convingut que s'ha recuperat prèviament per poder detectar només els enllaços que fan referència a notícies i que es troben al codi font de la pàgina web, eliminant tots els que facin referència a publicitat, altres seccions, contactes, ...
4. *Recuperar els descriptors que s'han d'utilitzar per la selecció*: el sistema ha d'anar a la Base de Dades i recuperar els descriptors que s'han d'utilitzar a la cerca del mitjà de comunicació que s'està extraient els enllaços de notícies.
5. *Descarregar el contingut dels enllaços trobats*: el sistema ha d'anar enllaç per enllaç obtingut al pas anterior i filtrar pels descriptors. Un cop es filtra poden donar-se dos vies.
6. *Verificació de coincidències*
   a. *No hi ha coincidència entre els descriptors i el contingut de la notícia*: es passa a l'enllaç següent i el sistema guarda l'enllaç per no consultar-ho en posteriors cerques.
   b. *Hi ha coincidència entre els descriptors i el contingut de la notícia*: ha d'analitzar el contingut de la notícia i localitzar les parts principals de les quals es composa. Per poder fer això, el sistema haurà d'utilitzar els patrons pertinents al mitjà de comunicació que estigui analitzant i utilitzar-los per localitzar les parts que desitja en cada moment.
7. *Emmagatzemar a la Base de Dades el resultat*: si el sistema ve del punt 5.a. llavors només emmagatzemarà a la taula DESCARTS l'enllaç corresponent. En canvi, si la línia del procés ve donada pel 5.b. llavors el sistema cal que emmagatzemi les parts de la notícia a la taula NOTICIA i les relacions pertinents a les taules MITJA_NOTICIA i DESCRIPTRO_NOTICIA per tenir enllaçada tota la informació per la seva posterior consulta. També, de la mateixa manera que succeeix amb els enllaços que no contenen la informació desitjada, un cop s'emmagatzema la informació l'enllaç de la notícia recaptada també s'afegirà a la taula DESCARTS per a que no es consulti en posteriors escombrats de la xarxa.

Aquest procés continua sent un procés iteratiu que l'administrador del sistema ha de decidir quan es llança, si ho fa manualment o automàticament, cada quan es vol fer escombrats dels mitjans de comunicació (cada hora, diaris, setmanals...). Aquestes decisions es prenen a mida que es va analitzant els mitjans de comunicació i es va veient la seva evolució i actualització de la informació que aporten, cada quan publiquen novetats o cada quan fan neteja de les notícies velles.

Tot seguit es pot veure el diagrama de Flux de Nivell 2 (*Figura 53*), que mostra tots els processos referents a la Gestió de Notícies i que no han canviat la seva manera d'interactuar amb l'usuari final de l'aplicació i que per tant és igual que l'esquema de la primera versió del projecte.

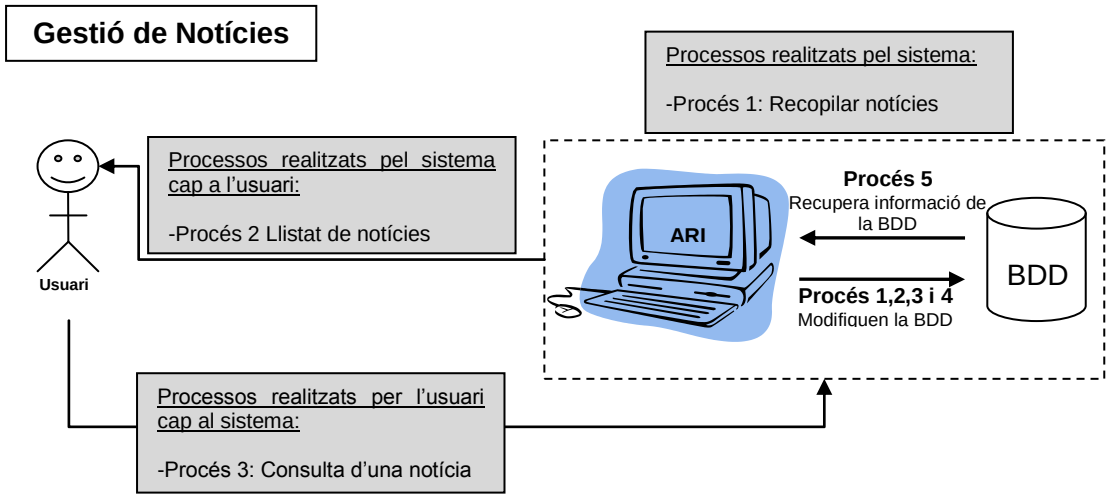

Figura 53: Diagrama de Flux de Dades de Nivell 2 – Gestió de Noticies (II)

9.6 Definició d'Interfícies d'Usuari (v.2)

Tot el que fa referència a aquests punts, es mantenen completament iguals que a la primera versió, i per tant, no es farà cap comentari al respecte, ja que a nivell d'interfície, l'usuari final no te constància de cap alteració del sistema, en el que a la utilització es refereix.

9.7 Definició de l'Arquitectura del Sistema (v.2)

9.7.1 Definició de Nivells d'Arquitectura (v.2)

Tal com succeeix amb les interfícies dels usuaris, l'arquitectura del sistema no pateix cap canvi substancial. Es segueix dividint en dos zones molt diferenciades, el servidor amb la Base de Dades i el programat per l'administrador del sistema, i per un altre costat està la Xarxa amb tota la informació que es vol recuperar i administrar segons convingui.

Així doncs, es passa a veure l'esquema de l'arquitectura (*Figura 54*):

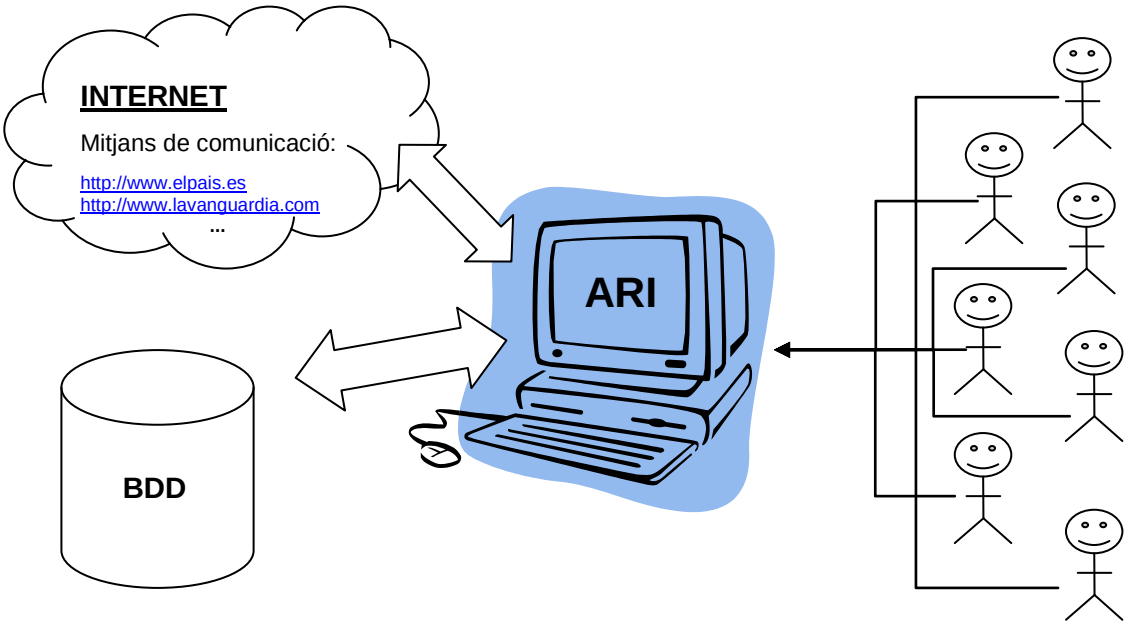

Figura 54: Esquema arquitectònic del sistema (II)

9.7.2 Especificació de l'Entorn Tecnològic (v.2)

Els requisits de l'entorn tecnològic tornen a ser iguals que els comentats al punt 7.7.2 d'aquesta documentació:

- L'administrador del sistema que munta el projecte necessita un ordinador o servidor amb connexió a Internet, amb un sistema de gestió de Base de Dades relacional (*MySQL*) i un servidor HTTP per poder gestionar (*Apache*)
- L'usuari només necessitarà un ordinador amb connexió a Internet.

9.8 Disseny de l'Arquitectura del Sistema (v.2)

9.8.1 Disseny de Mòduls del Sistema (v.2)

Al arribar un altre cop a aquesta altura del desenvolupament del projecte, els mòduls que forment el sistema continuen sent els mateixos, però amb canvis en la seva estructura de processos i aquest canvis són els que es passen a comentar.

A l'explicació dels mòduls de la primera versió, s'ha optat per la separació entre els mòduls des del punt de vista de l'usuari i des del punt de vista del sistema. Tal com ja s'ha comentat en algun dels punt anteriors, els processos que es miren des de la perspectiva de l'usuari no canvien i es mantenen tal qual s'han definit a la primera versió. En canvi, els processos dels mòduls des del punt de vista del sistema si que varien, sobretot dos: l'alta d'un nou mitjà de comunicació al sistema i la recaptació de notícies d'Internet.

A continuació es presenta el nou mòdul d'altes al sistema (*Figura 55*):

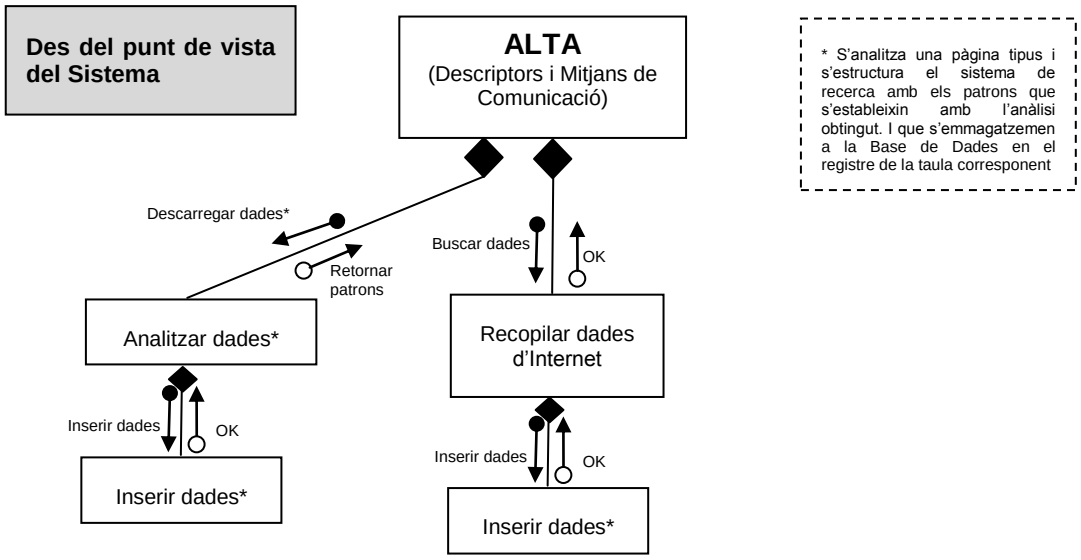

Figura 55: Diagrama d'estructura del mòdul d'altes (sistema) (II)

Tal com passa a la primera versió, en aquest cas, l'alta al sistema d'un mitjà de comunicació i una notícia seran recolzades per pseudocodis diferents. El que tenen en comú, és que prèviament, tal com s'observa a la *Figura 51*, s'haurà necessitat un pas previ per configurar la programació amb els patrons específics per cada mitjà de comunicació analitzat per poder realitzar posteriorment la cerca, tant d'enllaços com d'estructura de la notícia.

Pseudocodi d'alta d'un mitjà de comunicació:

```
Procediment ALTA_MITJA()
      dades←observarWeb()  //Administrador del Sistema
      afegir(dades)
Fprocediment
```

En aquest primer pseudocodi es pot observar com l'administrador del sistema haurà de fer el procés d'observar la web del mitjà de comunicació que es vol afegir al sistema i un cop s'han recopilat les dades necessàries (Nom, descripció i direcció web d'on es troba).

Fins aquí semblaria que no canvia res respecte a la primera versió, però a més a més de tot l'anterior, haurà de realitzar un anàlisi del codi font de la pàgina web per determinar els patrons que defineixen les estructures que necessitarà el procés posterior de recaptació de notícies, és a dir, el patró que trobarà els enllaços que siguin notícies, el patró corresponent al titular, el corresponent a l'entradeta i per últim, també, el corresponent al cos . Un cop es té donat d'alta al sistema ja es pot oferir a l'usuari com un mitjà més per als seus seguiments de premsa.

Tot seguit es passa a comentar el pseudocodi referent a l'obtenció de notícies d'Internet, les diferències entre la primera versió i la segona, aprofundir en les conseqüències dels canvis i les repercussions al sistema.

A continuació es detalla el pseudocodi d'alta d'una notícia:

```
Procediment ALTA_NOTICIA()
      Descriptors←recuperarDescriptors()
      Per cada mitja.actiu de la BDD fer
            patrons ← recuperarPatrons(mitja)
            dades←descarregarCodiFontWeb(mitja)
            enllaços←recuperarEnllaçosNotícies(dades,patro.enllaç)
            Per cada enllaç fer
                  noticia←recuperarNoticia(enllaç,patrons.noticia)
                  Si apareix(descriptors,noticia) llavors
                        afegirNoticia(noticia,dadesGenerals)
                  altrament
```

```
                        afegirDescart(enllaç)
                    fsi
                fper
            fper
        Fprocediment

Procediment recuperarEnllaçosNoticies(dades,patro)
    Enllaços ← trobarTotesLesCoincidencies(patro,dades)
    ← enllaços
Fprocediment

Procediment recuperarNoticia (enllaç,patrons)
    dades ← descarregarCofiFontWeb(enllaç)
    noticia.titular ← trobarTotesLesCoincidencies(patro,dades)
    noticia.titular ← trobarTotesLesCoincidencies(patro1,dades)
    noticia.entradeta ← trobarTotesLesCoincidencies(patro2,dades)
    noticia.cos ← trobarTotesLesCoincidencies(patro3,dades)
    ← noticia
Fprocediment

Procediment afegirNoticia(noticia,dadesGenerals)
    titular ← noticia.titular
    entradeta ← noticia.entradeta
    cos ← noticia.cos
    afegirALaBDD(titular,entradeta,cos,dadesGenerals)
Fprocediment
```

Amb el pseudocodi que aquí s'observa referent a l'alta d'una notícia, es pot veure els nous passos que haurà de realitzar el sistema per poder arribar a obtenir una notícia en aquesta segona versió del desenvolupament.

A continuació es passa a desenvolupar el detall de procediment del pseudocodi:

1. Es recuperen tots els descriptors pels quals s'haurà de filtrar les notícies.
2. Per cada mitjà actiu de la Base de Dades es recupera l'enllaç d'on es troba i el conjunt de patrons que estan relacionats amb ell.
3. Es descarrega el codi font per poder cercar els enllaços pertinents gràcies al patró de l'enllaç recuperat, mitjançant el procediment de *recuperarEnllaçosNoticies* explicat a la versió anterior i que es continua executant de la mateixa manera.
4. Per cada enllaç es descarrega el contingut i es localitzen les seves parts amb el procediment *recuperarNoticia*, que retorna un vector amb les diferents parts que es necessiten.
5. Es filtra per cada part els descriptors per veure si els troba:
    a. Si hi ha, s'afegeix la notícia (les seves parts i les dades generals) al sistema, emmagatzemant-ho correctament a la Base de Dades amb les corresponents relacions a les taules que ho necessiten (CLIENT_NOTICIA i DESCRIPTOR_NOTICIA).
    b. Si no hi és, només s'afegeix a la taula DESCART

Un cop s'han seguit tots els passos per tots els enllaços de mitjans de comunicació de la Base de Dades es trobarà actualitzat el sistema per a que l'usuari pugui visualitzar una nova remesa de notícies.

Aquest procés del mòdul d'altes és un procés que l'administrador determina quant es llença, així doncs, cal un estudi del món dels portals per determinar el temps òptim de cada quan cal programar una nova cerca pel sistema.

9.8.2 Exemple de cerca d'enllaços

A continuació es destacaran les diferències trobades a l'anàlisi del mitjà de comunicació El País i que han desencadenat els canvis comentats en els punts anteriors.

Figura 56: Portada de la web de ElPais.com

A la *Figura 56* es pot veure la portada web de ElPais.com a on es veuen les principals notícies reflectides. Si accedim al codi font de la portada de ElPais.com s'observa el següent:

```
</div>
                <!-- Inicio bloque A columna izq -->
<div class="bq_col1_a">

        <div class="mod_grafico_txt">


        <h2>
        <a href="/articulo/economia/deuda/familias/supone/115/renta/disponible/elpepueco/20070830elpepueco_6/Tes"  title="La subida del Euribor dispara la deu
</h2>

                <h4>

<a href="/articulo/economia/deuda/familias/supone/115/renta/disponible/elpepueco/20070830elpepueco_6/Tes"
</h4>
                <p>El endeudamiento familiar supone el 115% de su renta.- El &iacute;ndice de referencia hipotecario se sit&uacute;a en el nivel m&aacute;s al


                <div class="limpiar"></div>
                <!-- Fin relaciones -->
        </div>

</div>
<div class='filete_3'> </div>

                <!-- Inicio bloque A columna izq -->
<div class="bq_col1_b">


        <div class="mod_grafico_foto">

        <a href="/articulo/espana/Rosa/abandona/PSOE/sigue/siendo/socialista/elpepuesp/20070830elpepunac_5/Tes"  title=""Ha dejado de ser útil defender mis id
```

Figura 57: Codi font de la notícia de ElPais.com

Es pot observar a la *Figura 57* , el codi font de la portada de la pàgina web de ElPais.com, com el format de l'enllaç és diferent del comentat al punt 7.8.2.

El mateix passa amb la notícia i la seva estructura interna:

Figura 58: Notícia del ElPais.com

Aquesta notícia (*Figura 58*) porta associat el codi font següent:

```
      <div class="cabecera_noticia">
            <h2>  </h2>
            <h1>La deuda de las familias ya supone el 115% de su renta disponible</h1>
            <h3>El Euríbor sube por vigésimo tercer mes consecutivo y se sitúa en el nivel más alto desde diciembre de 2000 </h3>
            <div class="firma">
                  <p><strong>ELPAIS.com</strong> <em>- Madrid - </em>30/08/2007


                  </p>
        </div>
    <div class="limpiar"> </div>
  </div>

  <!-- ***** Contenido noticia ***** -->
      <div class="contenido_noticia">
    <!-- ***** Estructura_2col_1zq ***** -->
    <div class="estructura_2col_1zq">
      <div class="margen_n">
                        <div class="borde_sup"></div>
                        <!-- ***** Votos y comentarios ***** -->
                        <div class="votos">

<div id='votosC'>
      <div class="votos_estrellas">
<div class="votos_votar">
      <div class="votos_txt_vota">Vota</div>
         <div class="votos_estrella">
               <span id="bhs"><a class="aPs voto1" title="Sin interés" rel="nofollow" href='#%3Fctn%3DvotosC%26aP%3Dmodulo%253DEVN%2526params%253Did%25253D200
        </div><img src="/im/ico_separador_horizontal.gif" width="1" height="13" alt="" />
</div><div class="votos_resultados">Resultado <img src="/im/ico_vot_ok.gif" alt="Sin interés" title="Sin interés"><img src="/im/ico_vot_ok.gif" alt="Poco int
</div>
```

...

```
</div>
<p>Este incremento sin pausa ha ayudado, sin duda, a aumentar en los últimos años el endeudamiento familiar. Según un estudio de Caixa Catalunya, en 20
<div class="limpiar"></div>
```

Figura 59: Codi font de la notícia del ElPais.com

Aquesta *Figura 59* té una estructura diferent que la presentada a la primera versió del desenvolupament. Per tant, les expressions regulars configurades per fer la filtració tampoc serveixen. Això ha provocat aquesta nova versió i afrontar-la de manera que es configuri la Base de Dades de manera diferent.

La Base de Dades cal que emmagatzemi a la taula corresponent (on es guarden les dades del mitjà de comunicació) la configuració del patró de cada part del codi font de la pàgina web que es vol localitzar. És a dir, cal guardar per cada mitjà quatre patrons, el patró per localitzar els enllaços de les notícies, el patró que localitza el titular, el patró per l'entradeta i el patró pel cos.

9.9 Disseny físic de dades (v.2)

9.9.1 Disseny del Model Físic de Dades (v.2)

A continuació es presenta el model físic de les dades del sistema (*Figura 60*), que correspon totalment amb el de la primera versió. El que canvia en aquesta nova evolució és l'estructura de la taula MITJA que es passarà a comentar al punt següent de la memòria del projecte.

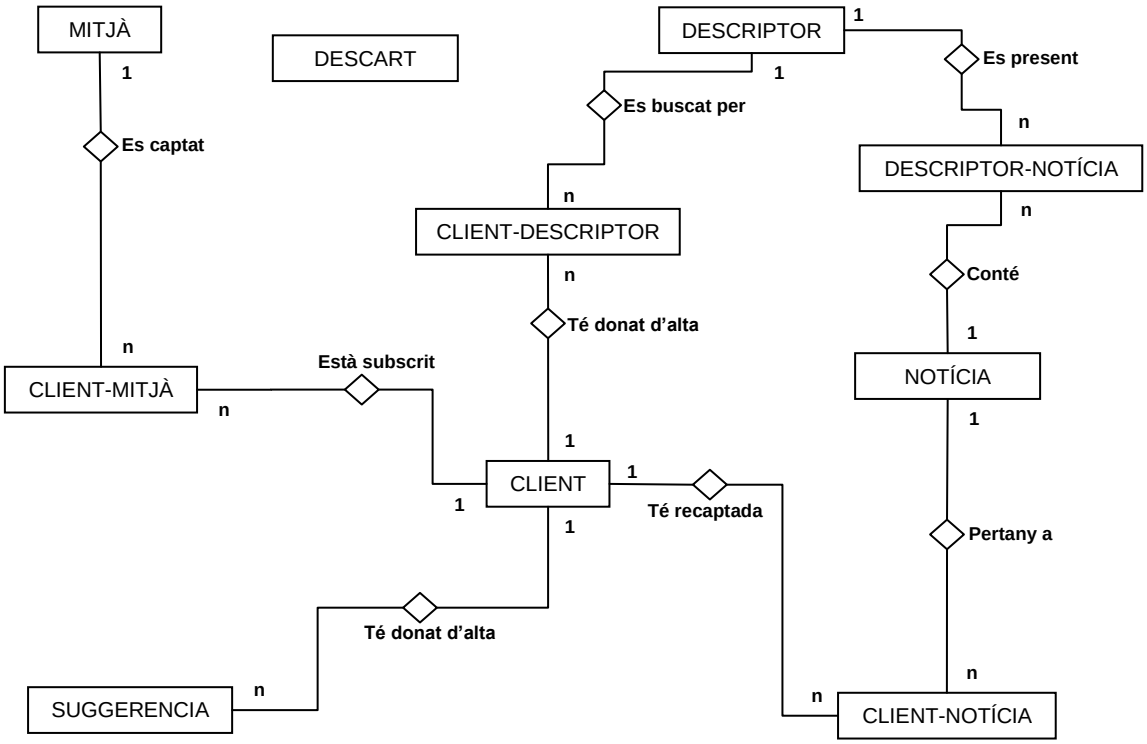

Figura 60: Diagrama d'Entitat-Relació

9.9.2 Descripció de les Taules (v.2)

Mitjà

Camp	Tipus	Descripció
MITJA_ID	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
MITJA_NOM	Varchar (100)	Nom del mitjà de comunicació
MITJA_DESCRIPCIO	Varchar (200)	Descripció del mitjà de comunicació
MITJA_URL	Varchar (150)	Enllaç a on es localitza el mitjà de comunicació
MITJA_ACTIU	Tinyint(1)	Número que indica si el mitjà està actiu per realitzar les cerques
MITJA_PATROENLLAS	Varchar (150)	Patró que identificarà els enllaços de notícies dins del codi font
MITJA_PATROTITULAR	Varchar (150)	Patró que identificarà el titular del codi font de la notícia
MITJA_PATROENTRADETA	Varchar (150)	Patró que identificarà l'entradeta del codi font de la notícia
MITJA_PATROCOS	Varchar (150)	Patró que identificarà el cos del codi font de la notícia

Taula 26: Mitjà (II)

Aquesta taula és la única que ha patit canvis de la primera versió a aquesta. Segueix emmagatzemant tota la informació necessària al respecte d'un mitjà de comunicació. D'aquesta manera, a part de les dades principals com el nom, la descripció i la direcció a on es troba localitzat el mitjà a Internet, també es seguirà necessitant el camp MITJA_ACTIU, un camp informatiu per a la programació que diu al sistema si aquest mitjà està habilitat per a la cerca o no i sobretot els quatre atributs que configuren els patrons de cerca que necessitarà el sistema per recuperar la informació de les diferents pàgines web de notícies descarregades del mitjà de comunicació en qüestió.

10 Construcció del Sistema d'Informació (v.2)

10.1 Execució de les proves unitàries i de les proves d'integració (v.2)

Al igual que succeeix amb les interfícies d'usuari del sistema, les proves que cal realitzar a la segona versió del desenvolupament són les mateixes que a la primera versió (punts 8.1 i 8.2 de la present memòria).

Tercera Versió del desenvolupament

11 Anàlisi i Disseny del Sistema d'Informació (v.3)

11.1 Definició del Sistema (v.3)

11.1.1 Determinació de l'Abast del Sistema (v.3)

Al finalitzar la segona versió del desenvolupament s'ha detectat com el procés de recaptació de notícies s'ha automatitzat molt més que a la primera versió, que només funcionava pel primer mitjà de comunicació analitzat.

De totes maneres, es van detectat petits errors de captures en mitjans de comunicació que s'havien analitzat al principi del procés. Això va provocar que es tingui que analitzar un altre cop el codi font dels mitjans de comunicació que fallen, per poder localitzar on radica l'error de funcionament. Després de fer aquest anàlisi s'ha pogut observar com els mitjans de comunicació han canviat la seva estructura interna del format HTML que utilitzen per la programació de les seves pàgines web.

Tot això ha provocat que es torni a dissenyar l'aplicació en una tercera versió per poder superar aquest nou inconvenient trobat. Aquesta nova versió del desenvolupament afectarà als mateixos punts crítics que s'han canviat de la primera a la segona versió, ja que els processos que es veuen afectats són els mateixos: l'alta de Mitjans de Comunicació i l'Obtenció de Notícies.

Així doncs, com a abast del sistema es segueix trobant els mateixos requeriments i objectius finals que a la primera y segona versió que ja s'han exposat en aquest document:

- *Obtenir notícies*: continua sent el fi bàsic del projecte, el responsable del canvi de versions del desenvolupament.
- *Gestionar els mitjans de comunicació*: igual en essència a la segona versió, amb petits canvis que s'exposaran al llarg del que queda de documentació.
- *Gestionar els descriptors:* exactament igual que a les versions anteriors.

Tot seguit es mostra el diagrama de context (*Figura 61*) a on es comencen a mostrar els canvis que necessitaran els processos per poder adaptar-se a la nova versió.

Diagrama de Context del Sistema (v.3)

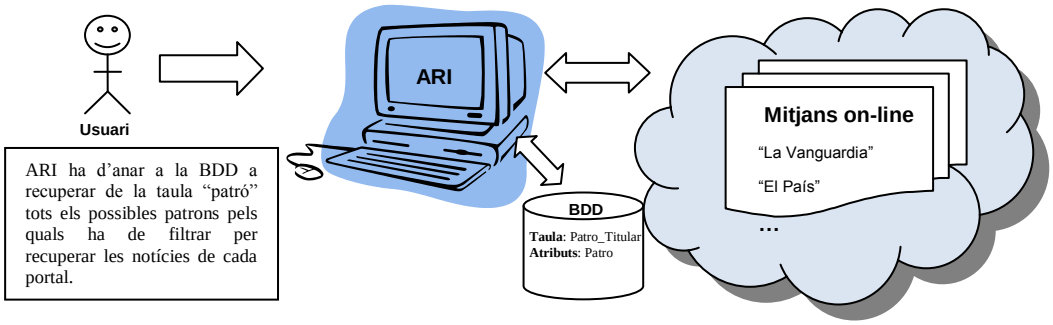

Figura 61: Diagrama de Context del Sistema (Versió 3)

11.1.2 Identificació d'Usuaris Participants i Finals (v.3)

Els usuaris participants en el procés de disseny, en la tercera versió del desenvolupament, passa a ser només un:

- Jo mateixa, que a aquestes altures del desenvolupament, el projecte ha passat a ser només responsabilitat meva.

11.2 Establiment de Requisits (v.3)

Aquesta nova versió dels requisits que ha d'afrontar el sistema, a part del superat de la primera a la segona versió (la NO homogeneïtat de l'entorn web), és la versalitat de codi, és a dir, la possibilitat de que un mitjà de comunicació canviï la seva estructura de codi font de les seves pàgines web, cada una quantitat de temps no conegut.

Tal com s'havia pensat en un principi el sistema, no es tenia present que entre diferents mitjans de comunicació poguessin tenir diferents estructures de codi font, aquest punt es va superar amb els canvis a la segona versió del desenvolupament, però no es va tenir en compte la possibilitat de que un mateix mitjà de comunicació pogués variar la seva estructura interna passat un cert temps. Això ha provocat que l'estructura interna del sistema desenvolupat a la segona versió hagi de patir canvis per poder solucionar aquest nou obstacle.

Així doncs, els requisits generals de l'aplicació continuen sent els mateixos però fent una revisió dels punts que s'han vist alterats ja a la segona versió:

- La transparència d'accés i gestió per a tenir organitzat tot el referent als mitjans de comunicació als quals s'està accedint durant el procés de recerca, el mateix al que ítems o descriptors es refereix, han d'estar introduïts a la Base de Dades.
- Una **Base de Dades que recolzi tot el procés**, les modificacions fetes a la segona versió (a la taula MITJA) no serviran i caldrà una revisió de l'estructura, que doni suport a la nova versió del desenvolupament.
- Una **algorísmica eficient**, i nova respecte a la segona versió en els punts referents a la captació d'un mitjà de comunicació nou i també respecte a la recaptació de notícies per la Xarxa, per a que sigui el més espontani i ràpid possible.

Un cop s'han tornat a comentar de nou els requisits bàsics del sistema es passarà a fer una re-especificació dels processos que han patit canvis, a causa dels nous requisits que s'han de solucionar respecte la primera i segona versió, per poder oferir un sistema a l'usuari estable i fiable.

11.2.1 La Gestió dels Mitjans de Comunicació (v.3)

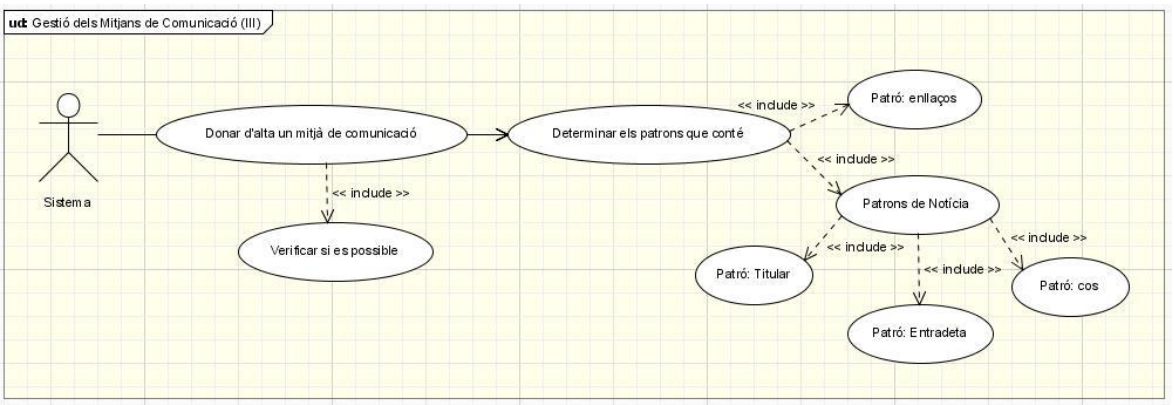

Figura 62: Cas d'ús – Gestió de Mitjans de Comunicació (III)

A aquesta *Figura 62* es pot observar la tercera versió del cas d'ús de Gestió de Mitjans de Comunicació. Aquesta nova versió pot semblar que és completament igual que l'exposat a la segona versió, pel que fa referència a la seva estructura, però cal comentar que el que canvia és els processos que porta a terme el sistema un cop a obtingut patrons. És a dir, com els emmagatzema a la Base de Dades i el posterior tractament que executarà el sistema per fer-los servir.

Pel que respecte a aquesta àrea de gestió en particular, si és mira des del punt de vista de l'usuari, aquesta continua sent exactament igual que a la segona versió, que tampoc canvia respecte al descrit a la primera versió del desenvolupament.

Per aquesta raó, a continuació es passa a descriure només el cas d'ús específic que ha patit els canvis d'una versió a l'altre: l'alta d'un mitjà de comunicació al sistema.

Cas d'ús	Alta d'un mitjà de comunicació (III)
Versió	1
Descripció	Donar d'alta un mitjà de comunicació.
Actors	Administrador del Sistema.
Precondició	El sistema està connectat a la xarxa.
Flux Principal	1. Descarregar el codi font de la pàgina principal del mitjà de comunicació 2. Analitzar el codi de la pàgina principal. c. Estructura dels enllaços. d. Confecció del patró per localitzar-los. 3. Descarregar el codi d'una notícia del mitjà de comunicació. c. Estructura d'una notícia. d. Confeccionar els patrons de les parts de la notícia. 4. **Guardar els patrons al corresponent a les taules de patrons.** **a. Patró de l'enllaç a la taula: ENLLAÇ** **b. Patró del tituar a la taula: TITULAR** **c. Patró de l'entradeta a la taula: ENTRADETA** **d. Patró del cos a la taula: COS** 5. **A les noves taules que relacionen els registres de la taula MITJA amb les taules de patrons, actualitzar la relació.** 6. **Estructurar el codi del sistema per a que agafi cada configuració de patrons.**
Postcondició	Codi confeccionat de manera que cerqui mitjançant els patrons.

Taula 27: Fitxa – Alta d'un mitjà de comunicació (III)

A la *Taula 27* es pot veure la fitxa corresponent al cas d'ús d'alta d'un mitjà de comunicació. Aquesta nova versió té present els problemes apareguts al llarg de la transició entre la primera versió i la segona i té present els nous inconvenients sorgits al llarg del desenvolupament de la segona versió.

Així doncs, es troba com encara que en aparença l'estructura del procés que recolza el cas d'ús és la mateixa, en el fons, els procediments que hi ha darrera han canviat substancialment la seva manera de funcionar. Han aparegut noves taules a nivell físic i una nova manera d'entendre l'anàlisi i recaptació d'informació d'una pàgina web d'un mitjà de comunicació.

Ara, l'administrador del sistema, cada cop que doni d'alta un nou mitjà a la Base de Dades o revisi informació d'un mitjà ja existent, caldrà que faci una gestió més acurada de les noves dades recuperades, ja que un mitjà de comunicació no tindrà només associat un patró de cada tipus possible (enllaç, titular, entradeta o cos) sinó que existeix la possibilitat de que tingui varis per cada patró. Per aquesta raó apareixen al sistema noves taules i nous procediments per portar terme tan el cas d'ús que aquí es tracta com el referent a l'obtenció de notícies futures.

La gestió de les noves taules es portarà a terme cada cop que es realitzi una nova inserció d'un mitjà de comunicació o revisió d'un ja existent. D'aquesta manera, l'administrador del sistema haurà de controlar si el patró que troba es nou o no, modificar la seva probabilitat dins del sistema pel mitjà que s'està analitzant i posteriorment en el procés d'obtenció de noticies caldrà configurar els processos de tal manera que recuperin sempre la millor opció que ofereix el sistema, tenint en compte sempre la resta per si un cas no funciona la primera opció i així iterativament fins trobar l'opció apropiada o avisar al sistema d'un buit en el procés, però això s'explicarà en detall en el punt següent de la documentació.

En definitiva, l'alta d'un nou mitjà de comunicació es passa a utilitzar tant per un nou mitjà de comunicació al sistema, com per la revisió d'un ja existent i haurà de tenir en compte un seguit de paràmetres:

- Localització de les dades generals (en cas de nova alta al sistema)
- Localització dels patrons actuals de les pàgines web relacionades amb el mitjà
- Actualització de la Base de Dades amb el contingut dels patrons relacionats amb el mitjà.
- Actualització de les noves probabilitats d'obtenció de notícies segons el patró utilitzat a cada moment.

11.2.2 L'obtenció de Notícies (v.3)

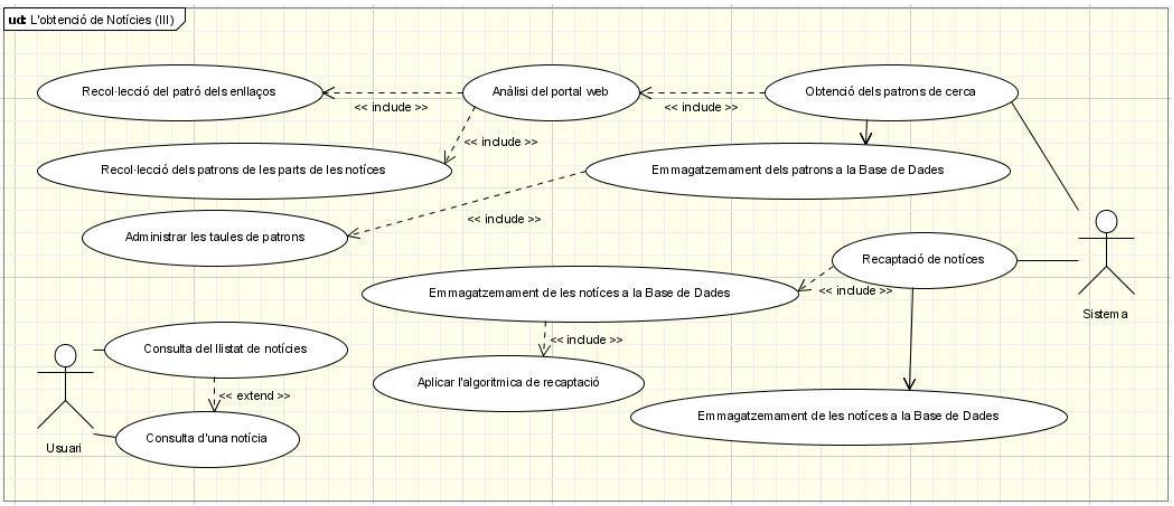

Figura 63: Cas d'ús – L'obtenció de Notícies (III)

A la *Figura 63* que aquí s'observa es pot veure com el procés de captació de noves notícies. Aquest procés sembla similar al procés de captació de la segona versió del desenvolupament. A continuació es presenta la fitxa corresponent a l'àrea del cas d'ús a on s'han patit els canvis que s'han de portar a terme.

Cas d'ús	**Recaptació de notícies (III)**
Versió	1
Descripció	Recol·lectar les notícies de les webs dels mitjans de comunicació.
Actors	Sistema.
Precondició	El sistema està connectat a la xarxa.

Flux Principal	1. **Recuperar de la Base de Dades els patrons referents al mitjà de comunicació que es vol extreure la informació.**
	  a. **Extreure per cada cas el patró amb probabilitat més alta.**
	2. Descarregar el codi font de la pàgina principal del mitjà de comunicació
	3. Extreure els enllaços corresponent a notícies **mitjançant el patró d'enllaços que s'ha recuperat.**
	  a. **Si es troba els enllaços desitjats passar al següent punt**
	  b. **Si no es localitzen els patrons, recuperar el següent patró de la Base de Dades amb més probabilitat i provar-ho**
	  i. **Si es troba el patró que funciona, s'actualitzen les probabilitats a la Base de Dades**
	  ii. **Si no es troba, avisar al sistema per a que revisi el mitjà de comunicació. Es para el procés de captació pel mitjà actual.**
	4. Descarregar el codi d'una notícia del mitjà de comunicació.
	5. Filtrar si s'ha d'emmagatzemar (segons apareixen els descriptors, o no)
	6. Si apareixen els descriptors, extreure la informació necessària de la notícia **mitjançant els demés patrons extrets de la Base de Dades (repetir el mateix procés que amb els patrons d'enllaços).**
	  a. Extreure el titular
	  b. Extreure l'entradeta
	  c. Extreure el text
	7. Emmagatzemar al sistema amb les relacions pertinents.

Postcondició	Descriptor esborrat del sistema

Taula 28: Fitxa – Recaptació de Notícies (III)

Tal com passa amb el pas de la primera a la segona versió del desenvolupament, el procés d'obtenció de notícies ha patit un canvi en la seva algorítmica. Al modificar-se l'estructura de la Base de Dades tant a nivell dels atributs de la taula MITJA, com del sorgiment de noves taules, fa que la manera de treballar del sistema es modifiqui significativament. Al llarg dels següents punts de la documentació es deixarà patent la nova forma de treballar que tenen els processos per poder tenir present i solucionar tots els obstacles que es plantegen en els requeriments de la nova versió.

11.3 Identificació de Subsistemes d'Anàlisi (v.3)

11.3.1 Identificació i Definició de Subsistemes (v.3)

A continuació es presenten les noves característiques que formen part dels subsistemes que ja s'han trobat a la primera i segona versió del desenvolupament.

De la mateixa manera que a les versions anteriors, els subsistemes que formen part del projecte continuen sent els mateixos. En el cas del subsistema de Descriptors, les semblances arriben als processos, les funcionalitats... . En canvi, els subsistemes de Mitjans de Comunicació i Notícies, encara que es segueixen la mateixa estructura que a les versions anteriors, el seu contingut a nivell de processos canvia per poder donar servei als nous requeriments de la versió actual que es vol desenvolupar.

Tot seguit es mostra el diagrama de Flux de Dades general del sistema i posteriorment es comentaran les àrees específiques que pateixen els canvis de la nova versió.

Diagrama de Flux de Dades de nivell 1 (v.3)

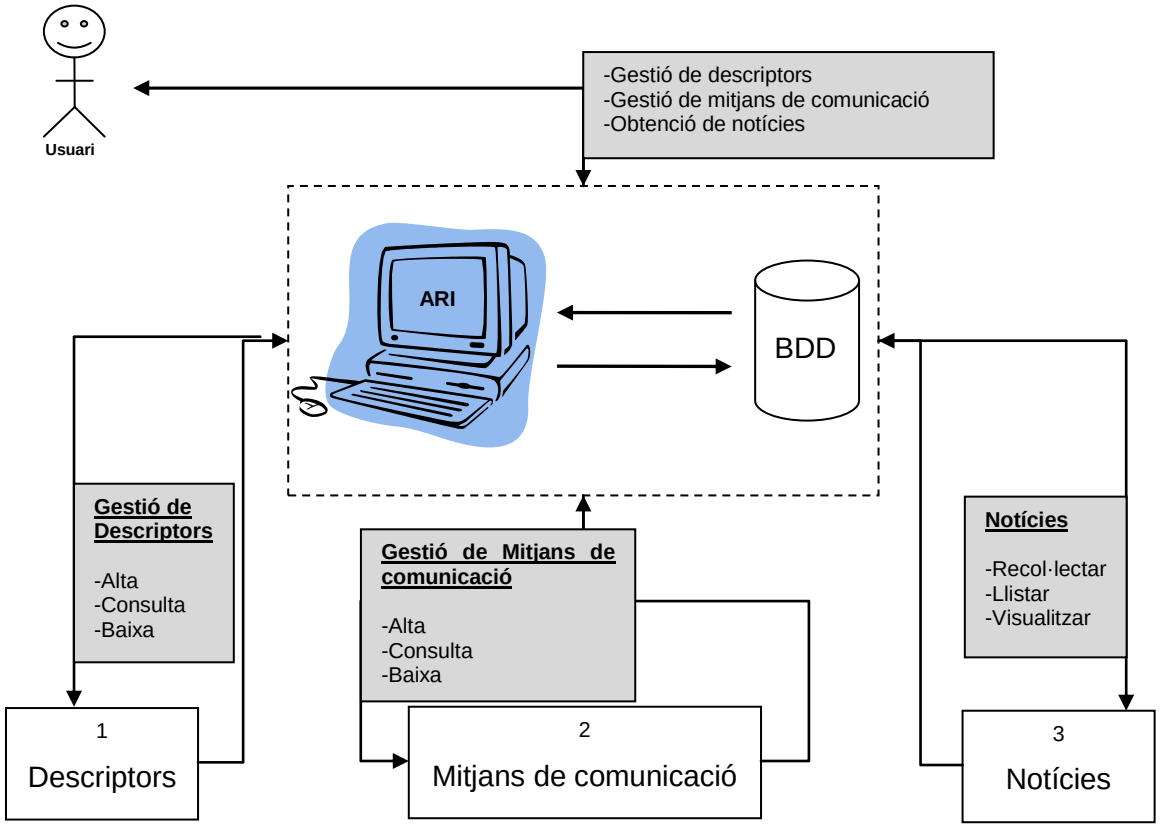

Figura 64: Diagrama de Flux de Dades de Nivell 1 (v.3)

Així doncs, a l'observar la *Figura 64*, a l'igual que passa amb la segona versió del desenvolupament, es veu com el diagrama és el mateix que a la primera versió. Això passa perquè el que es canvia no és el flux de dades en sí, no és el traspàs d'informació dins del sistema ni les dades que finalment arriben a l'usuari, sinó que són els processos que canvien la seva forma de treballar dins del sistema.

Per fer un recordatori del que engloba cada subsistema es pot anar al punt 7.3.1 de la documentació.

11.4 Elaboració del Model de Dades (v.3)

11.4.1 Model Lògic de Dades Demanat (v.3)

Un cop es té clar que els subsistemes identificats segueixen sent els mateixos i que el que canvia són els procediments que executen, es repassa el següent punt de la primera i segona versió: el model lògic de dades.

A aquest punt passa el mateix que el comentat a l'apartat anterior (la identificació dels subsistemes), en aparença tot segueix igual, el diagrama de classes d'aquesta versió es gairebé igual que el de la primera i la segona versió, només afegint una classe més que es pot visualitzar a la *Figura 65* següent:

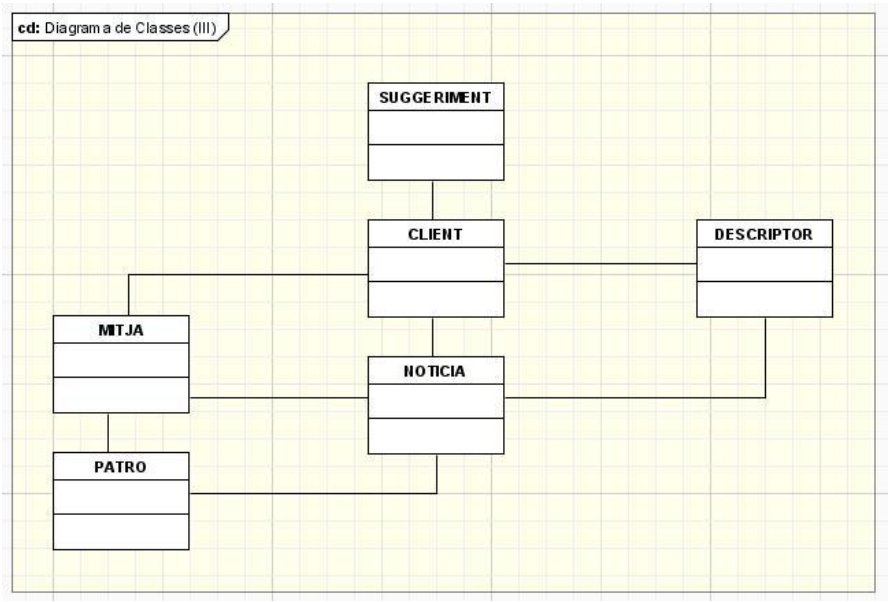

Figura 65: Diagrama de Classes (III)

El diagrama de classes és igual als exposats a la primera i segona versió, sumant-li una classe nova: PATRO que dona suport per les noves necessitats que ha de solucionar l'aplicació. De totes maneres, el pes central del sistema continua girant al volta de les classes NOTICIA i CLIENT, que distribueixen les accions bàsiques que cal portar a terme per donar servei a l'usuari.

Al igual que passa amb la segona versió, encara que el diagrama continua sent molt semblant al presentat a la primera i segona versió, els mètodes que conté cada classe, així com la forma d'implementar-los és diferent a cada versió presentada.

A continuació es presentaran els canvis que es donen en els processos del sistema que aporten al sistema les funcionalitats tal com les necessita l'usuari per obtenir els resultats desitjats a l'utilitzar l'aplicació.

11.5 Elaboració del Model de Processos (v.3)

Tal com s'ha fet a la segona versió del desenvolupament, a continuació s'expliquen els canvis que s'han donat en els processos dels diferents subsistemes de l'aplicació.

Per tenir una visió més clara i concisa de les diferents àrees del model de processos es continua amb la forma de treball que a les versions anteriorment comentades. S'ha dividit el model en els diferents subsistemes, per poder aportar una explicació més detallada dels processos que conté cadascun.

Cal comentar que el model de processos referent a la Gestió de Descriptors no canvia d'una versió a l'altre i que, per tant, a continuació no se'n farà esment.

11.5.1 Gestió dels Mitjans de Comunicació (v.3)

Les funcionalitats que ha de solucionar aquest subsistema continuen sent les especificades al punt 9.5.1 de la primera versió del desenvolupament. Fins i tot, els processos que es declaren des del punt de vista de l'usuari són completament iguals. En canvi, els processos des del punt de vista del sistema si que pateixen canvis, sobretot el referent a donar d'alta un mitjà de comunicació, tal com s'ha anat explicant en la introducció d'aquesta nova versió.

Tot seguit es presenta el detall d'aquest procés que es troba dins d'aquest subsistema:

Procés 1: donar d'alta un Mitjà de Comunicació.

En aquesta versió es troba que l'administrador del sistema ha de donar d'alta el mitjà de comunicació a la Base de Dades per a que l'aplicació pugui fer la cerca, al igual que succeeix amb la primera versió i la segona. D'igual manera, aquest procés el segueix tinent que portar a terme l'administrador del sistema, que haurà d'incloure a la Base de Dades la informació necessària del nou mitjà de comunicació que es vulgui donar d'alta: nom, descripció, enllaç i si està actiu o no. La diferència radica que ara caldrà realitzar l'anàlisi de l'HTML, de com està estructurat el codi font a la web i extreure del seu anàlisi els patrons declarats a segona versió. La diferència amb la segona versió és el que l'administrador del sistema ha de portar a terme un cop ha localitzat aquests patrons. Cal que s'emmagatzemin a les noves taules i fer els càlculs necessaris per establir les probabilitats (explicades en el punt de Disseny de Mòduls del Sistema, punt 11.8.1) que necessitarà el procés d'obtenció de notícies que s'explicarà en el següent punt.

Tot seguit, a la *Figura 66* es mostra el global de processos que engloben aquest subsistema i que només ha patit canvis el primer.

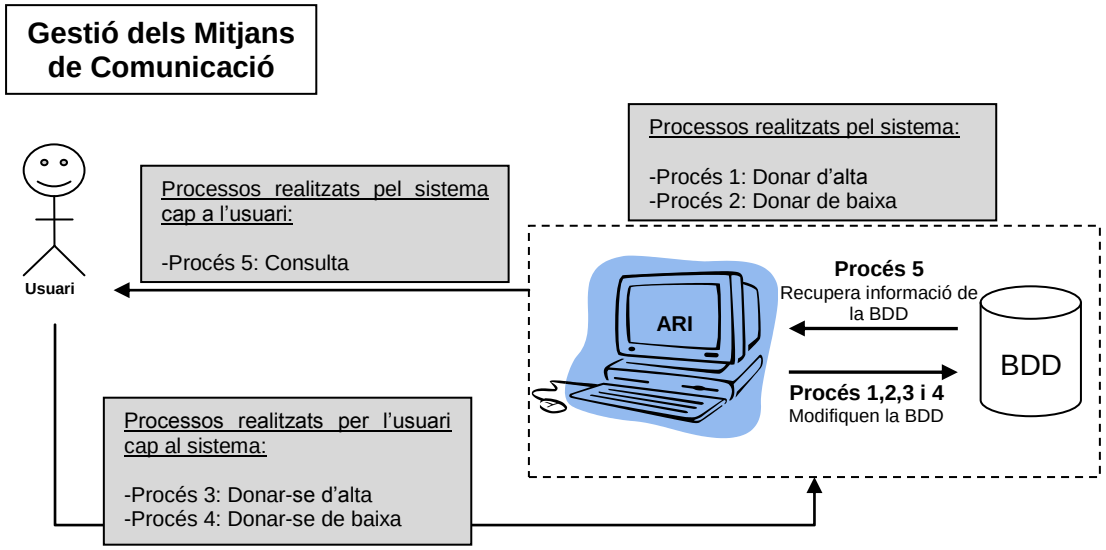

Figura 66: Diagrama de Flux de Dades de Nivell 2 – Gestió de Mitjans de Comunicació (III)

11.5.2 L'Obtenció de Notícies

Els processos d'aquests subsistema són els encarregats de recaptar la informació de la xarxa, igual que a les versions anteriors del desenvolupament del sistema. El que aquesta nova versió aporta és un canvi radical en la seva algorítmica, en el que el procés de recaptació es refereix.

Els demés processos que s'han comentat a la primera versió i que tampoc es comenten a la segona, són els que no canvien en res la seva manera de funcionar, a cap nivell, ni algorítmic ni funcional. De totes maneres, el diagrama de Flux de Dades de Nivell 2 (*Figura 59*) representa el global dels processos que es troben en aquest subsistema.

A continuació es passa a comentar el detall del procés de recopilar notícies, que és el que es veu modificat per culpa dels nous requeriments de la tercera versió.

Procés 1: recopilar Notícies d'Internet

El procés de recopilar notícies de la Xarxa continua sent el més important de tot el projecte i el que fa que es revisin les versions un cop que es consideren finalitzades.

Aquest procés està dirigit pel sistema i és l'encarregat d'anar per les pàgines web dels mitjans de comunicació i buscar les notícies que corresponguin amb els paràmetres establerts a la Base de Dades en cada moment.

A la primera versió es va pensar que les pàgines web serien homogènies en la seva estructura HTML, a la segona versió es va millorar el sistema afegint els patrons adaptats a cada mitjà de comunicació i com que al finalitzar aquesta versió es van seguir localitzant fallades, va provocar el re-anàlisi de les estructures dels mitjans de comunicació que ja s'havien analitzat per arribar a la conclusió que es necessitaria una nova versió que s'adaptés als nous descobriments: les pàgines web d'un mitjà poden canviar amb el temps: no és fix i per tant no pot aplicar-se l'esquema de la Base de Dades establert a la segona versió del desenvolupament.

Aquest descobriment ha provocat canvis a diversos nivells: per començar es procedeix a comentar les passes que en aquesta nova versió el sistema ha de portar a terme per recuperar les notícies:

1. *Recuperar els patrons de la Base de Dades associats a cada Mitjà de Comunicació*: cal que el sistema primerament vagi a la Base de Dades a recuperar la informació sobre els patrons que estan associats al codi font del mitjà de comunicació del qual es vol recuperar la informació. Aquest patrons es recuperaran de la taula que associa el patró amb el mitjà de comunicació i aquest serà escollit segons la major probabilitat que se l'hagi donat, per ser més viable a l'hora de donar un bon resultat a la cerca.
2. *Recuperar enllaços dels Mitjans de Comunicació*: cal que el sistema vagi a la Base de Dades i recuperi la direcció web, l'enllaç, a on es pot trobar el contingut de cada mitjà.
    a. Si s'han pogut recuperar el enllaços es passa al punt 3.
    b. Si no s'han pogut recuperar els enllaços es passa a buscar per la Base de Dades un patró de més a menys probabilitat d'encert per mirar si encaixa, si no encaixa cap dels propis del mitjà, es passarà a buscar per tots els de la taula del patró corresponent.
        i. Si encaixa un patró del mitjà de comunicació, es recalculen les probabilitats de tots els patrons associats al mitjà, a part de recuperar els enllaços i seguir amb el fil normal d'execució.
        ii. Si encaixa un patró que no està associat al mitjà de comunicació, se li associa i es recalculen les probabilitats, i es segueix amb el fil d'execució recuperant els enllaços i seguint amb el procediment.
        iii. Si no encaixa cap patró s'avisa a l'administrador del sistema que cal tornar a analitzar l'estructura de la pàgina web del mitjà de comunicació
3. *Descarregar el contingut de l'enllaç*: cal que el sistema descarregui la informació de la pàgina web i trobi tots els enllaços, aplicant el patró convingut que s'ha recuperat prèviament per poder detectar només els enllaços que fan referència a notícies i que es troben al codi font de la pàgina web, eliminant tots els que facin referència a publicitat, altres seccions, contactes, ...
4. *Recuperar els descriptors que s'han d'utilitzar per a la selecció*: el sistema ha d'anar a la Base de Dades i recuperar els descriptors que s'han d'utilitzar a la cerca del mitjà de comunicació que s'està extraient els enllaços de notícies.

5. *Descarregar el contingut dels enllaços trobats*: el sistema ha d'anar enllaç per enllaç obtingut al pas anterior i filtrar pels descriptors. Un cop es filtra poden donar-se dos vies.
6. *Verificació de coincidències*
    a. *No hi ha coincidència entre els descriptors i el contingut de la notícia*: es passa a l'enllaç següent i el sistema guarda l'enllaç per no consultar-ho en posteriors cerques.
    b. *Hi ha coincidència entre els descriptors i el contingut de la notícia*: ha d'analitzar el contingut de la notícia i localitzar les parts principals de les quals es composa. Per poder fer això, el sistema haurà d'utilitzar els patrons pertinents al mitjà de comunicació que estigui analitzant i utilitzar-los per localitzar les parts que desitja en cada moment.
        i. Si encaixa el patró més probable recuperat de la Base de Dades inicialment es continua pel pas 7.
        ii. Si encaixa un dels patrons següents en probabilitat, es continua el procés i es recalculen les probabilitats per posteriors recaptacions.
        iii. Si encaixa un dels altres patrons, caldrà associar-lo al mitjà de comunicació, recalcular les probabilitats i continuar amb el procés del punt 7.
        iv. Si no encaixa cap patró, caldrà avisar a l'administrador del sistema de que cal tornar a analitzar el codi font de la pàgina web del mitjà de comunicació que s'està cercant.
7. *Emmagatzemar a la Base de Dades el resultat*: si el sistema ve del punt 6.a. llavors només emmagatzemarà a la taula DESCARTS l'enllaç corresponent. En canvi, si la línia del procés ve donada pel 6.b. (una de les tres primeres opcions) llavors el sistema cal que emmagatzemi les parts de la notícia a la taula NOTICIA i les relacions pertinents a les taules MITJA_NOTICIA i DESCRIPTRO_NOTICIA per tenir enllaçada tota la informació per la seva posterior consulta. També, de la mateixa manera que succeeix amb els enllaços que no contenen la informació desitjada, un cop s'emmagatzema la informació l'enllaç de la notícia recaptada també s'afegirà a la taula DESCARTS per a que no es consulti en posteriors escombrats de la xarxa.

Cal tenir en compte que en tot moment que s'utilitza un patró per fer la cerca, tan sigui dels enllaços de notícies, com després de les parts que està formada aquesta, pot ser que cap dels patrons que es tingui associats al mitjà de comunicació funcioni, per això cal tenir preparat el sistema per a que avisi a l'administrador de que en el mitjà de comunicació en qüestió ha ocorregut una excepció i que cal tractar-la.

Aquest procés continua sent un procés iteratiu que l'administrador del sistema ha de decidir quan es llança, si ho fa manualment o automàticament, cada quan es vol fer escombrats dels mitjans de comunicació (cada hora, diaris, setmanals...). Aquestes decisions es prenen a mida que es va analitzant els mitjans de comunicació i es va veient la seva evolució i actualització de la informació que aporten, cada quan publiquen novetats o cada quan fan neteja de les notícies velles, també, en aquesta nova versió cal tenir en compte cada quan els programadors canvien l'estructura interna de les seves pàgines web, per poder avançar-se als possibles errors de recaptació.

Tot seguit es pot veure el diagrama de Flux de Nivell 2 (*Figura 67*), que mostra tots els processos referents a la Gestió de Notícies i que no han canviat la seva manera d'interactuar amb l'usuari final de l'aplicació i que per tant és igual que l'esquema de la primera versió del projecte.

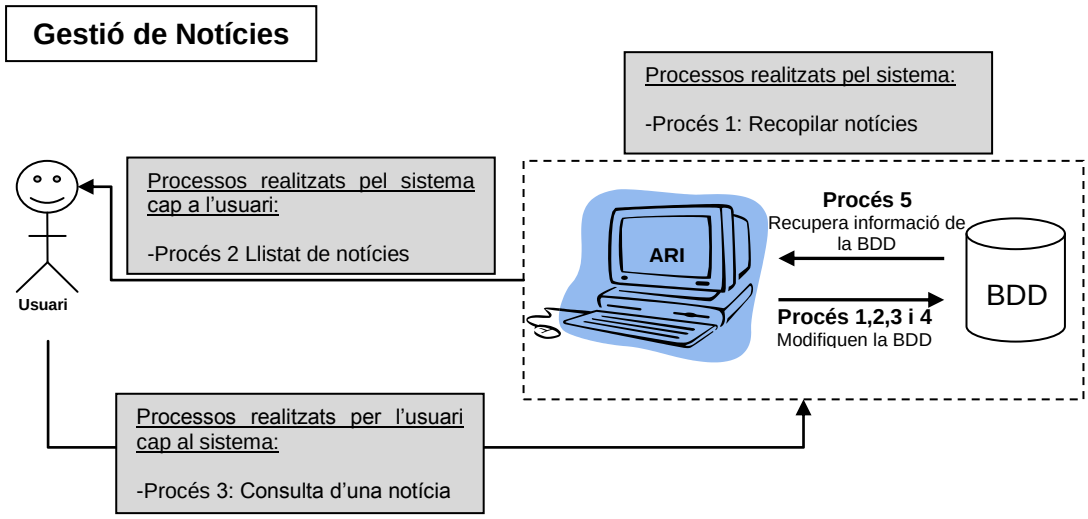

Figura 67: Diagrama de Flux de Dades de Nivell 2 – L'obtenció de Notícies (III)

11.6 Definició d'Interfícies d'Usuari (v.3)

11.6.1 Especificació de Principis Generals de la Interfície (v.3)

Tal com passa amb la segona versió, tot el que afecta a l'usuari final de l'aplicació es continua respectant igual que es troba a la primera versió del desenvolupament i que per tant aquest punt de la documentació és exactament igual al esmentat a l'apartat 7.7 .

11.7 Definició de l'Arquitectura del Sistema (v.3)

11.7.1 Definició de Nivells d'Arquitectura (v.3)

Tal com succeeix amb les interfícies dels usuaris, l'arquitectura del sistema no pateix cap canvi substancial. Es segueix dividint en dos zones molt diferenciades, el servidor amb la Base de Dades i el programat per l'administrador del sistema i per un altre costat està la Xarxa amb tota la informació que es vol recuperar i administrar segons convingui.

Així doncs, es passa a veure l'esquema de l'arquitectura (*Figura 68*):

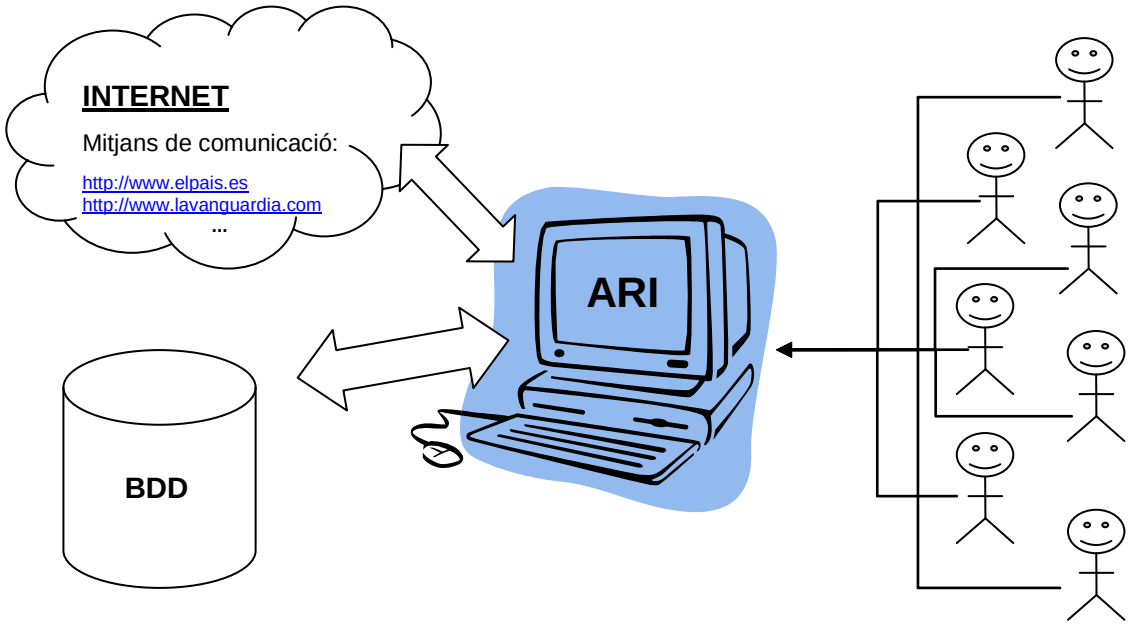

Figura 68: Esquema arquitectònic del sistema (III)

11.7.2 Especificació de l'Entorn Tecnològic (v.3)

Els requisits de l'entorn tecnològic tornen a ser iguals que els comentats al punt 7.7.2 d'aquesta documentació:

- L'administrador del sistema que munta el projecte necessita un ordinador o servidor amb connexió a Internet, amb un sistema de gestió de Base de Dades relacional (*MySQL*) i un servidor HTTP per poder gestionar (*Apache*)
- L'usuari només necessitarà un ordinador amb connexió a Internet.

11.8 Disseny de l'Arquitectura del Sistema (v.3)

11.8.1 Disseny de Mòduls del Sistema (v.3)

En aquest punt del desenvolupament, a la tercera iteració que es dóna per solucionar els obstacles trobats, es passa a comentar els mòduls que formen part del sistema. Aquest continuen sent els mateixos, però, tal com succeeix en la transició de la primera a la segona versió, es troben canvis en l'estructura des processos i són aquestes modificacions les que es passen a comentar a continuació.

Si s'observen les explicacions de les dues primeres versions, es pot veure com a aquesta es continua optant per la mateixa separació dels mòduls: des del punt de vista de l'usuari i des del punt de vista del sistema. De la mateixa manera que passa a les anteriors versions, els processos des del punt de vista de l'usuari no canvien i es mantenen tal qual s'han definit a la primera versió del desenvolupament, per un altre costat, a l'igual que passa amb la segona versió, els processos dels mòduls des del punt de vista del sistema si que varien, sobretot dos: l'alta d'un mitjà de comunicació al sistema i la recaptació de notícies d'Internet.

A continuació es presenta el nou mòdul d'altes al sistema (*Figura 69*):

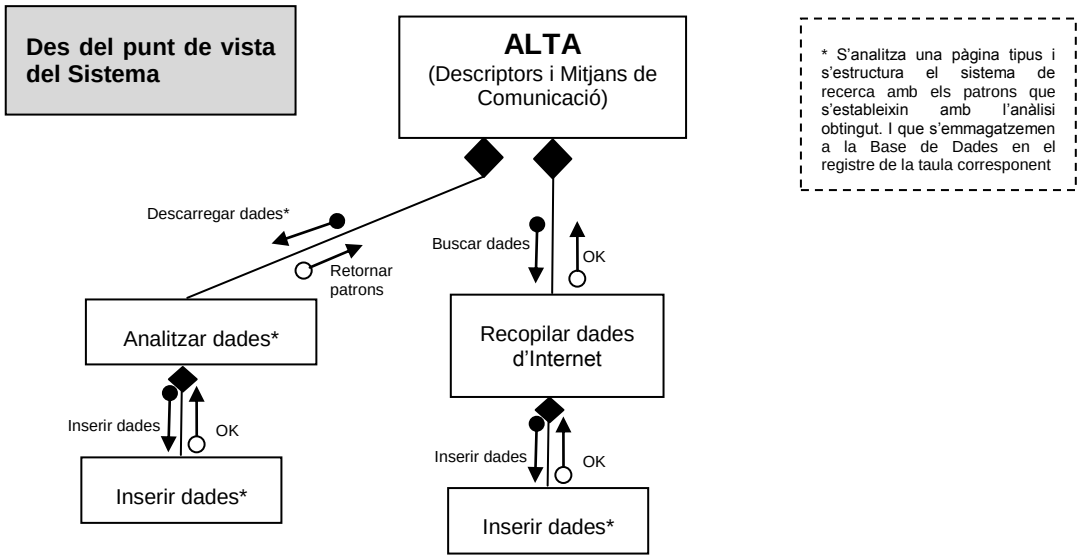

Figura 69: Diagrama d'estructura del mòdul d'altes (v.3)

Al igual que passa amb la segona versió, el pseudocodi que recolzarà els processos d'aquest mòdul (*Figura 69*) d'altes seran diferents per tractar els casos d'altes de mitjans de comunicació i captures de notícies.

Tots dos casos d'altes cal que vagin a la Xarxa a descarregar-se el contingut del mitjà de comunicació que es vol tractar i un cop es té la informació passar a tractar-la. Arribats a aquest punt, els camins dels dos casos es separen i es passen a tractar individualment.

Pseudocodi d'alta d'un mitjà de comunicació:

```
Procediment ALTA_MITJA()
dades←observarWeb()    //Administrador del Sistema
afegir(dades)
Fprocediment
```

```
Procediment afegir(dades)
  patrons←recuperarPatrons(dades)
    Per cada patro fer
       analitzar(patro)
    Fper
Fprocediment
```

```
Procediment analitzar(patro)
  Si existeix(patro) llavors
    RecalcularProbabilitat(patro)
  altrament
    afegir(patro)
  FSi
Fprocediment
```

Al pseudocodi referent a l'alta d'un mitjà de comunicació o de la seva revisió, és bastant diferent en la seva línia algorítmica que el de la segona versió.

Potser, en un primer moment pot semblar que la filosofia és la mateixa, ja que continua descarregant-se la informació del mitjà de comunicació i localitzant les dades que són necessàries pel sistema. La diferència radica en el tractament que se'n fa d'aquestes dades un cop adquirides.

El sistema ha de veure si ja les té incorporades, s'està fent referència als patrons trobats que possibiliten la cerca al mitjà de comunicació, i si ja les té incorporades cal que es recalculin els pesos de les probabilitats de tots els possibles patrons que té associat el mitjà de comunicació, perquè el posterior procés de captació pugui agafar la millor possibilitat a la primera passada d'anàlisi de les dades.

Tot seguit es passa a comentar el pseudocodi referent a l'obtenció de notícies d'Internet, les diferències entre la primera versió i la segona, aprofundir en les conseqüències dels canvis i les repercussions al sistema.

Pseudocodi d'alta d'una notícia:

```
Procediment ALTA_NOTICIA()
  Descriptors←recuperarDescriptors()
  Per cada mitja.actiu de la BDD fer
    patrons ← recuperarPatronsMesProbables(mitja)
    dades←descarregarCodiFontWeb()
    enllaços←recuperarEnllaçosNotícies(dades,patro.enllaç)
    si (enllaços.tamany >0) llavors
      per cada enllaç fer
        noticia←recuperarNoticia(enllaç,patrons.noticia)
        Si (apareix(descriptors,noticia)) llavors
          si (partsNoticies.plenes) llavors
            afegir(partsNotices,enllaç)
          altrament
            coincideix←provarAltresPatronsNoticia(enllaç,mitja)
            si NO coincideix llavors
              avisarSistemaDeCanvis(mitja)
            fsi
          fsi
        altrament
            afegirDescart(enllaç)
        fsi
      fper
    altrament
      coincideix←provarAltresPatronsEnllaços(enllaç,mitja)
      si NO coincideix llavors
        avisarSistemaDeCanvis(mitja)
      fsi
    fsi
  fper
Fprocediment
```

Pseudocodi referent al procediment executat al procés principal, aquest busca els demés patrons possibles per trobar les parts de la notícia en el contingut passat per paràmetre.

```
Procediment provarAltresPatronsNoticia(enllaç,mitja)
  Mentre NO trobat i i<mitja.QuantitatPatronsAssociats fer
    Patrons←recuperarSegüentPatronsNoticiesProvable(i,mitja)
    Si funcionen(patrons,enllaç) llavors
      afegirNoticia2(patrons,enllaç)
      recalcularProbabilitats(patrons,enllaç)
      trobat← cert
    altrament
      i←i+1
    fsi
  fmentre
  i←0
  demesPatrons←recuperarDemesPatronsNoticia(mitja)
  Mentre NO trobat i i<demesPatrons.Quantitat fer
    Si funcionen(patrons,enllaç) llavors
      afegirNoticia2(patrons,enllaç)
      afegirProbabilitats(patrons,enllaç)
      trobat← cert
    altrament
      i←i+1
    fsi
  fmentre
←trobat
fprocediment
```

A continuació es pot veure el pseudocodi referent a un dels procediments executats al procés principal, aquest busca els demés patrons possibles per trobar els enllaços a notícies en el contingut passat per paràmetre.

```
Procediment provarAltresPatronsEnllaços(enllac,mitja)
  Mentre NO trobat fer
    Patro←recuperarSegüentPatroEnllaçProvable(i,mitja)
    Si funciona(patro,enllaç) llavors
      buscarNoticies(patro,enllaç)
      recalcularProbabilitats(mitja,patro)
      trobat← cert
    altrament
      i←i+1
    fsi
  fmentre
  i←0
  demesPatro←recuperarDemesPatronsEnllaç(mitja)
  Mentre NO trobat i i<demesPatrons.Quantitat fer
    Si funcionen(patrons,enllaç) llavors
      afegirNoticia2(patrons,enllaç)
      afegirProbabilitats(patrons,enllaç)
      trobat← cert
    altrament
      i←i+1
    fsi
  fmentre
←trobat
fprocediment
```

A aquest pseudocodi referent a la captació de notícies de la Xarxa, es pot observar els nous passos que el sistema haurà de realitzar per poder arribar a obtenir una col·lecció de notícies en aquesta tercera versió del desenvolupament.

A continuació es detalla línia a línia del procés principal i dels procediments que conté:

1. Es recuperen tots els descriptors pels quals s'haurà de filtrar les notícies.
2. Per cada mitjà actiu de la Base de Dades es recupera l'enllaç d'on es troba i els patrons MÉS PROBABLES que estan relacionats amb ell, és a dir, es busca dins de la relació entre el mitjà i les taules de patrons aquell que la seva probabilitat d'encert sigui més alta.
3. Es descarrega el codi font per poder cercar els enllaços pertinents gràcies al patró de l'enllaç recuperat.
    a. Si es troben enllaços amb el patró més probable, continuar amb el punt 4
    b. Si no es troben enllaços cal fer un procés iteratiu anomenat *provarAltresPatronsEnllaços*
        i. Va recuperant de la Base de Dades els patrons per ordre, de més probable a menys, fins que:
            1. Si coincideix algun, s'insereixen les corresponents dades a la Base de Dades i es recalculen els pesos de les probabilitats .
            2. Si no coincideix cap s'avisa a l'administrador del sistema de que cal tornar a analitzar l'estructura de les pàgines web del mitjà de comunicació que s'està tractant.
        ii. Si s'esgoten els possibles patrons i no es troba solució cal avisar al sistema.
4. Per cada enllaç es descarrega el contingut i es filtra si s'hi troba algun dels descriptors.

a. Si hi ha, es busquen les diferents parts de la notícia amb els patrons més probables recuperats prèviament.
    i. Si coincideix el patró, s'insereix a la Base de Dades la informació.
    ii. Si no coincideixen els patrons més probables repetir com al punt 3.b.
    iii. Si no queden més patrons al sistema que provar, avisar a l'administrador del sistema de que cal tornar a estudiar l'estructura de les pàgines web del mitjà de comunicació que s'està tractant.
b. Si no hi és, només s'afegeix a la taula DESCART

Un cop s'han seguit tots els passos per tots els enllaços de mitjans de comunicació de la Base de Dades es trobarà actualitzat el sistema per a que l'usuari pugui visualitzar una nova remesa de notícies.

En el cas de que el procés hagi invocat un avís, es passarà a fer un repàs del mitjà de comunicació en qüestió per tal d'actualitzar les seves dades. També, en cas de no ser el principal patró (el més probable) el bo per l'obtenció de la notícia, en el procés de proba d'altres patrons caldrà que l'administrador del sistema realitzi el procés de recalcular les probabilitats per a posteriors cerques.

Aquest procés del mòdul d'altes és un procés que l'administrador determina quant es llença, així doncs, cal un estudi del món dels portals per determinar el temps òptim de cada quan cal programar una nova cerca pel sistema.

11.8.2 Càlcul de les probabilitats

Un cop arribat aquest punt, i vist la gran varietat de patrons associats que pot arribar a tenir un mitjà de comunicació, s'ha decidit que la versió tres del desenvolupament incorpori un sistema de pesos (probabilitats) lligats als patrons.

Per poder instaurar un sistema de probabilitats s'ha hagut de pensar una fórmula que donés el rendiment desitjat. Els resultats esperats es voldran obtenir en funció de dos paràmetres:

- *Duració*: el sumatori de tots els períodes de temps que el patró ha estat actiu en el procés de cerca pel contingut del mitjà de comunicació. Es vol que la probabilitat sigui proporcional a la duració.
- *Antiguitat*: el temps que fa que es va donar d'alta al sistema .Es vol que els patrons més antics tinguin una probabilitat més baixa.

Aquest dos paràmetres han sigut els escollits per configurar la fórmula de la probabilitat que tot seguit es mostra:

$$Probabilitat = (Duració/Antiguitat)/Norma$$

Es pot observar com a la fórmula apareix un paràmetre no comentat, aquest representa el factor responsable de que el resultat que s'obtingui estigui normalitzat (entre 0 i 1). A continuació és mostra aquesta fórmula:

$$Norma = \sum_{Patrons} Duració/Antiguitat$$

Amb aquest paràmetres es vol aconseguir que el patró actual, la seva probabilitat sigui 1/Norma, o sigui, la màxima. A partir del moment que es detecta que el patró actual ja no fa el servei esperat, es quan la fórmula dissenyada pren el seu protagonisme. Un exemple:

> *El patró que localitza els enllaços de ElPais.com ha estat donat d'alta al sistema el dia 1 del mes. El sistema guarda que la seva Antiguitat i la seva Duració són les mateixes, el que fa que a l'aplicar la fórmula en tot moment el resultat sigui el màxim (1/Norma) que és l'esperat.*

Un cop el sistema detecta el dia 30 del mes, que el patró actual no és vàlid, llança la fórmula per recalcular la probabilitat del patró en qüestió. Aquesta detectarà que la duració del patró és de 29 dies i que l'antiguitat del mateix de 30, així doncs el resultat sempre serà estrictament menor que 1.

Així doncs la fórmula aportarà al sistema una configuració de pesos que donarà la possibilitat d'optimitzar el procés de cerca de notícies, superant els obstacles que s'ha trobat a l'acabar la segona versió del desenvolupament.

11.9 Disseny físic de dades (v.3)

11.9.1 Disseny del Model Físic de Dades (v.3)

Tot seguit es presenta el model físic de les dades del sistema (*Figura 70*), que respecte al model presentat a la primera i segona versió, afegeix taules al sistema i modifica la taula MITJA de tal manera que els seus atributs són iguals, però el sentit que tenen canvia.

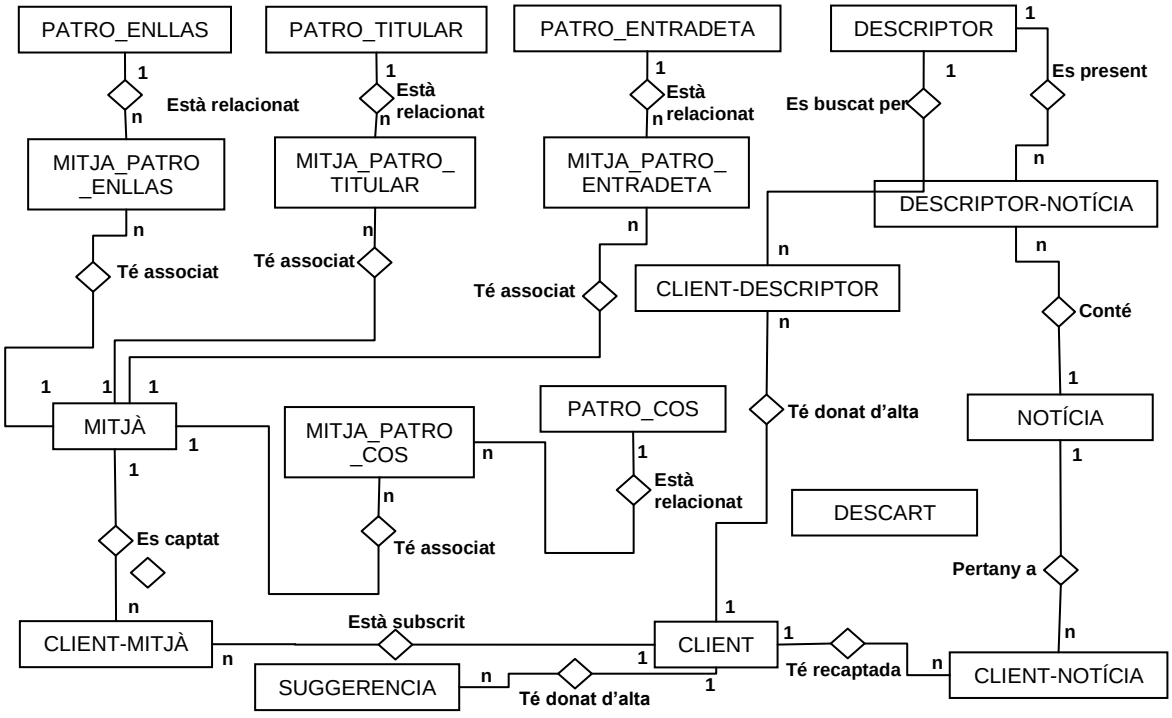

Figura 70: Diagrama d'Entitat-Relació (v.3)

11.9.2 Descripció de les Taules (v.3)

A continuació es mostren les taules noves i la modificació dels atributs de la taula MITJA.

Mitjà

Camp	Tipus	Descripció
MITJA_ID	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
MITJA_NOM	Varchar (100)	Nom del mitjà de comunicació
MITJA_DESCRIPCIO	Varchar (200)	Descripció del mitjà de comunicació
MITJA_URL	Varchar (150)	Enllaç a on es localitza el mitjà de comunicació
MITJA_ACTIU	Tinyint(1)	Número que indica si el mitjà està actiu per realitzar les cerques
MITJA_PATROENLLAS	Varchar (150)	Patró que identificarà els enllaços de notícies dins del codi font
MITJA_PATROTITULAR	Varchar (150)	Patró MES PROBABLE que identificarà el titular de la notícia
MITJA_PATROENTRADETA	Varchar (150)	Patró MES PROBABLE que identificarà l'entradeta la notícia
MITJA_PATROCOS	Varchar (150)	Patró MES PROBABLE que identificarà el cos de la notícia

Taula 29: Mitja (III)

Aquesta taula continua emmagatzemant tota la informació necessària al respecte d'un mitjà de comunicació. De fet, tots els camps de la taula són exactament iguals que a la versió anterior, amb els patrons que aquest té associats. La diferència vindrà donada pel sentit que tenen aquests atributs que contenen els patrons. Aquests atributs contindran el patró més probable per l'obtenció del resultat desitjat. Aquest atribut s'actualitzarà al mateix temps que les taules que relacionen les taules *patrons* amb la taula mitjà.

Les següents quatre taules: PATRO_ENLLAS, PATRO_TITULAR, PATRO_ENTRADETA i PATRO_COS estan dissenyades per emmagatzemar a la Base de Dades tots els patrons trobats al llarg de l'anàlisi de tots els mitjans de comunicació i estan directament relacionades amb els mitjans de comunicació gràcies a les quatre taules: MITJA_PATRO_ENLLAS, MITJA_PATRO_TITULAR, MITJA_PATRO_ENTRADETA i MITJA_PATRO_COS.

D'aquesta manera es tindrà un banc de patrons que s'utilitzarà per realitzar les cerques de notícies. Aquestes cerques tindran varies fases segons a quin nivell es trobin les coincidències entre el que es vol buscar i els patrons que s'utilitzen per fer-ho.

Així doncs, primer el cercador anirà a buscar a la taula MITJA, a on està emmagatzemat el patró amb més alt nivell de probabilitat de coincidència, posteriorment buscarà a les taules que relacionen el mitjà amb el tipus de patrons els següents més probables i en última instància es buscarà a tota la taula del patró corresponent coincidències amb els patrons d'altres mitjans.

En el cas de que es trobi una coincidència a l'últim nivell, o sigui, que es doni per bo un patró associat a un mitjà de comunicació que no sigui l'actual, el sistema haurà d'actualitzar la taula MITJA_PATRO_X i afegir la nova relació amb la probabilitat calculada associada.

En el últim cas de no trobar a cap nivell un patró que doni coincidències a la cerca, cal que el sistema actualitzi les dades sobre el mitjà de comunicació corresponent i actualitzar la informació a les taules de la Base de Dades pertinent.

Patro_Enllas

Camp	Tipus	Descripció
PATRO_ENLLAS_ID	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
PATRO_ENLLAS	Varchar (100)	Expressió Regular

Taula 30: Patro_enllas

Patro_Titular

Camp	Tipus	Descripció
PATRO_TITULAR _ID	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
PATRO_TITULAR	Varchar (100)	Expressió Regular

Taula 31: Patro_titular

Patro_Entradeta

Camp	Tipus	Descripció
PATRO_ENTRADETA _ID	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
PATRO_ENTRADETA	Varchar (100)	Expressió Regular

Taula 32: Patro_entradeta

Patro_Cos

Camp	Tipus	Descripció
PATRO_ENTRADETA _ID	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
PATRO_ENTRADETA	Varchar (100)	Expressió Regular

Taula 33: Patro_cos

Mitja_Patro_Enllas

Camp	Tipus	Descripció

MITJA_PATRO_ENLLAS_ID_MITJA	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
MITJA_PATRO_ENLLAS_ID_ENLLAS	Varchar (100)	Nom del mitjà de comunicació
MITJA_PATRO_ENLLAS_PROBABILITAT	Smallint (5)	Probabilitat d'èxit associada al mitjà de comunicació
MITJA_PATRO_ENLLAS_ANTIGUITAT	Smallint(5)	Antiguitat, en dies, que porta l'enllaç al sistema
MITJA_PATRO_ENLLAS_DURACIO	Smallint(5)	Duració, en dies, que ha estat bo el patró per la captació

Taula 34: Mitja_Patro_enllas

Mitja_Patro_Titular

Camp	Tipus	Descripció
MITJA_PATRO_TITULAR_ID_MITJA	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
MITJA_PATRO_TITULAR_ID_TITULAR	Varchar (100)	Nom del mitjà de comunicació
MITJA_PATRO_TITULAR_PROBABILITAT	Smallint (5)	Probabilitat d'èxit associada al mitjà de comunicació
MITJA_PATRO_TITULAR_ ANTIGUITAT	Smallint(5)	Antiguitat, en dies, que porta l'enllaç al sistema
MITJA_PATRO_TITULAR_ DURACIO	Smallint(5)	Duració, en dies, que ha estat bo el patró per la captació

Taula 35: Mitja_Patro_titular

Mitja_Patro_Entradeta

Camp	Tipus	Descripció
MITJA_PATRO_ENTRADETA_ID_MITJA	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
MITJA_PATRO_ENTRADETA_ID_ENTRADETA	Varchar (100)	Nom del mitjà de comunicació
MITJA_PATRO_ENTRADETA_PROBABILITAT	Smallint (5)	Probabilitat d'èxit associada al mitjà de comunicació
MITJA_PATRO_ENTRADETA_ANTIGUITAT	Smallint (5)	Antiguitat, en dies, que porta l'enllaç al sistema
MITJA_PATRO_ENTRADETA_DURACIO	Smallint (5)	Duració, en dies, que ha estat bo el patró per la captació

Taula 36: Mitja_Patro_entradeta

Mitja_Patro_Cos

Camp	Tipus	Descripció
MITJA_PATRO_COS_ID_MITJA	Smallint (5)	Identificador de la taula, està configurat amb autoincrement
MITJA_PATRO_COS_ID_COS	Varchar (100)	Nom del mitjà de comunicació
MITJA_PATRO_COS_PROBABILITAT	Smallint (5)	Probabilitat d'èxit associada al mitjà de comunicació
MITJA_PATRO_COS_ANTIGUITAT	Smallint (5)	Antiguitat, en dies, que porta l'enllaç al sistema
MITJA_PATRO_COS_DURACIO	Smallint (5)	Duració, en dies, que ha estat bo el patró per la captació

Taula 37: Mitja_Patro_cos

12 Construcció del Sistema d'Informació (v.3)

12.1 Execució de les proves unitàries i de les proves d'integració (v.3)

Al igual que succeeix amb les interfícies d'usuari del sistema, les proves que cal realitzar a la segona versió del desenvolupament són les mateixes que a la primera versió (punts 8.1 i 8.2 de la present memòria).

13 Ampliacions i millores

En aquest punt del desenvolupament de la tercera versió es dóna per finalitzat el projecte final de carrera, ja que es considera que s'ha arribat a aconseguir els objectius que s'especificaven al inici de la documentació.

Aquest projecte es deixa en una fase de proves on s'ha d'anar veient l'efectivitat de l'algorisme probabilístic implementat i valorar les diferents opcions de disseny del mateix, si en un futur es veiessin caigudes importants del sistema. Per això cal tenir en compte que a dia d'avui es deixa un sistema en funcionament per mitjans de comunicació en el que no ha hagut canvis en la seva estructura interna des de que s'ha posat en funcionament el projecte, de tal manera no s'ha pogut valorar la capacitat de l'algorisme probabilístic dissenyat per donar solucions als possibles problemes al llarg de la recaptació de notícies.

Si es centra la visió a en les funcionalitats que l'aplicació dóna a l'usuari, es veu una línia clara d'ampliació, noves possibilitats a oferir a l'usuari, com són la possibilitat d'unes interfícies més gestionables, a mode de carpetes que facilitessin l'administració tant dels descriptors com de les notícies, podent així repartir la informació d'una manera forma i intuïtiva. Això oferiria a l'usuari la possibilitat de tenir un magatzem de notícies administrat tal com ell mateix decideixi.

Per un altre costat, com a possible millora, durant tot el projecte es diu que l'administrador del sistema ha d'anar realitzant inspeccions periòdiques a l'estructura web de cada mitjà de comunicació, per poder detectar els patrons que sorgeixen per poder captar les diferents informacions que es volen emmagatzemar a la Base de Dades. Aquest procés manual es podria automatitzar gràcies a mètodes d'Intel·ligència Artificial que detectessin automàticament els patrons al text, es necessitarien tècniques de reconeixement de text i algorismes preparats per la confecció automàtica d'expressions regulars.

Totes dues vessant, tant la millora del sistema com la posterior ampliació, impliquen l'estudi d'un món nou i que només s'ha deixat entreveure al llarg del desenvolupament d'aquest projecte: la Intel·ligència Artificial. Aquesta ciència s'intuïa que seria necessària en algun moment del desenvolupament donat el caire desconegut del problema al qual es volia donar solució al inici del projecte. La aplicació d'Intel·ligència Artificial al desenvolupament futur del projecte obriria els horitzons de les possibilitats d'un cercador de notícies a Internet.

14 Conclusions

Com a introducció es vol destacar l'evolució del projecte a nivell de responsabilitats que s'han tingut al llarg del seu desenvolupament, posteriorment s'exposaran les implicacions que ha tingut el desenvolupament d'una aplicació d'aquest estil al llarg de tot el temps que ha durat el Projecte Final de Carrera, tant a nivell acadèmic com personals .

Fa un any i mig, quan es va començar a desenvolupar el projecte, havia entrat a treballar a una empresa que es dedica a donar servei de notícies als seus clients. Aquest servei, a l'origen de l'empresa només es feia de pont entre els clients i una altra empresa que sí realitzava les cerques a la Xarxa. Però al veure oportunitat de negoci, van decidir implementar la seva pròpia aranya cercadora i no haver de dependre d'un proveïdor de notícies. En aquest punt vaig arribar a l'empresa, just presa la decisió de començar el disseny d'aquesta aranya. Vaig poder participar en el procés de selecció de la plataforma, el llenguatge de programació, la Base de Dades i el seu corresponent gestor. I el responsable de l'empresa em va donar total llibertat per investigar sobre el món d'Internet i els portals dels mitjans de comunicació. Aquesta responsabilitat va suposar un gran incentiu per portar a terme el desenvolupament del projecte i quan a mitjans de la segona versió es va decidir no acabar d'implementar-ho, per les raons ja comentades durant la documentació, vaig demanar poder utilitzar el desenvolupat fins al moment com a Projecte Final de Carrera.

Després de l'acceptació per part de l'empresa de poder prosseguir amb el projecte fora de l'empresa vaig tenir la primera toma de contacte amb el director (Gustavo Patow) que ha acabat supervisant l'execució del projecte. Li vaig exposar la naturalesa del projecte i en quin punt estava i vam decidir continuar-ho per presentar-ho com a Projecte Final de Carrera d'Enginyeria Informàtica.

Al llarg de tot el temps que s'ha necessitat pel desenvolupament del projecte les reunions amb el director han estat periòdiques i molt orientatives per poder portar per bons camins l'execució de tot el sistema que es volia implementar. L'ajuda en temes de planificació de tasques i prioritats han estat essencials per poder organitzar correctament el treball i portar-ho fins al punt que aquí es deixa.

Per un altre costat, les implicacions a nivell acadèmic que ha provocat el desenvolupament del projecte han desembocat en l'obtenció de nous coneixements, tant sobre el llenguatge PHP utilitzat a la programació, com en tècniques d'optimització d'algorismes de captació i gestió de taules a Base de Dades. També s'ha començat a entreveure la necessitat d'adquirir nous coneixements en tècniques d'Intel·ligència Artificial com són els sistemes d'agents i multiagents.

Com a conclusió final destacar que a nivell personal aquest projecte a suposat un gran salt a nivell de coordinació de projectes ja que m'ha donat la possibilitat de veure la meva capacitat organitzativa, quins són els meus punts forts de treball i en quines àrees necessito adquirir més pràctica.

Agraïments

Com a punt final de la documentació crec important reservar aquest punt per agrair a tothom implicat al meu voltant l'ajuda i suport que m'han donat al llarg d'any i mig de desenvolupament del projecte i que han fet la tasca una mica mes lleugera.

Primerament agrair a l'empresa i als responsables originaris del projecte poder continuar amb ell i fer-ho meu després de descartar-lo com projecte empresarial. La possibilitat d'utilitzar tot el desenvolupat a llarg dels mesos que vaig treballar i el seu suport en moments clau del procés han fet que el projecte hagi pogut arribar a aquest punt.

Agrair al director del projecte, Gustavo Patow, que ha supervisat tot el desenvolupament un cop vaig començar a treballar sense l'empresa. Al llarg de tot aquest temps ha donat la seva visió i aportat la seva experiència, fent del desenvolupament del projecte quelcom més fàcil.

Per últim, agrair especialment a la meva família i amics. A la família per donar-me suport moral i als amics per ajudar-me en moments d'estancament i aportar la seva visió crítica com a companys d'estudis.

A tothom, moltíssimes gràcies!

15 Referències

15.1 Suport a la documentació

[1] Notícia (Agost 2007)
http://es.wikipedia.org/wiki/Noticia

[2] RSS (Agost 2007)
http://es.wikipedia.org/wiki/RSS

[3] Press-Clipping (Agost 2007)
http://www.zonalibre.org/blog/Carpanta/archives/095948.html

[4] cUrl (Agost 2007)
http://www.desarrolloweb.com/faq/que-es-curl.html
http://www.php.net/curl

[5] HTML (Agost 2007)
http://es.wikipedia.org/wiki/HTML

[6] Expressions Regulars (Agost 2007)
http://es.wikipedia.org/wiki/Expresi%C3%B3n_regular

[7] mySQL (Agost 2007)
http://es.wikipedia.org/wiki/MySQL

[8] PHP (Agost 2007)
http://es.wikipedia.org/wiki/Php

[9] Extreme Programing (Agost 2007)
http://es.wikipedia.org/wiki/Programaci%C3%B3n_Extrema

[10] Esquema per les Fitxes de Cas d'ús (Agost 2007)
http://ima.udg.edu/Docencia/3105IG0008/

[_] Gran Diccionari de la Llengua Catalana (Agost 2007)
http://www.diccionari.cat/

[_] Diccionario Castellano de la Real Academia Española (Agost 2007)
http://www.rae.es/

15.2 Suport a la programació

[11] CSS (Agost 2007)
http://www.w3schools.com/css/default.asp

[12] cUrl (Agost 2007)
http://www.desarrolloweb.com/articulos/utilizar-curl-para-copiar-imagen-al-disco.html

[13] PHP (Agost 2007)
http://es2.php.net/manual/es/function.preg-match-all.php